شناسایی سرقت ادبی مبتنی بر الگوریتم ژنتیک و برچسب‌گذاری نقش‌معنایی در مقالات علمی

رضوان یعقوبی[1]،حسن ختنلو[2]

۱کارشناس ارشد، دانشگاه آزاد اسلامی، واحد ملایر Rezvaneyaghobi2050@gmail.com

۲دانشیارگروه مهندسی کامپیوتر،دانشگاه بوعلی سینا، همدان

تاریخ دریافت: ۹۲/۱۰/۲ تاریخ پذیرش: ۹۴/۵/۱۷

چکیده

امروزه با پیشرفت روز افزون اینترنت و گسترش مقالات برخط دستبردهای علمی راحت‌تر شده است. سرقت ادبی استفاده دوباره یا کپی کردن متنی بدون ارجاع به نویسنده‌ی اصلی است. سرقت علمی یا تقلب در مدارس و دانشگاهها می‌تواند به عنوان یک فاکتور محرک برای معلمان، دانش آموزان، دانشجویان و اساتید به حساب آید. اگر سرقت علمی و ادبی به درستی شناسایی نشود، متقلبان و سارقان می‌توانند به نتایجی برسند که مستحق آن نیستند. در این مقاله روشی جهت شناسایی سرقت ادبی بر مبنای برچسب‌گذاری نقش معنایی و الگوریتم ژنتیک ارائه می‌شود. روش پیشنهادی بر روی متون انگلیسی عمل پردازش را انجام می‌دهد. نتایج آزمایش بر روی مجموعه داده‌های PAN-PC-09 نشان می‌دهد که روش پیشنهادی، مقدار پارامترهای ارزیابی مانند Recall، Precision و F-measure را نسبت به روش‌های قبلی ارائه شده در زمینه شناسایی سرقت ادبی بهبود می‌دهد.

کلیدواژه

سرقت ادبی،سرقت معنایی، الگوریتم ژنتیک، برچسب گذاری نقش معنایی،محاسبه‌ی شباهت

مقدمه

شناسایی شباهت متون یکی از شاخه‌های متن‌کاوی[1] است که کاربرد آن در شناسایی سرقت ادبی[2]است. سرقت ادبی به معنای گرفتن (ایده، اسناد، کد، عکس و ...) از دیگران و انتصاب (ایده، اسناد، عکس و ...) به نام خود بدون ذکر منبع و مرجع است [۱].

دسترسی آسان به وب و پایگاه داده‌های بزرگ و به طور کلی ارتباطات ازراه دور باعث شده که سرقت ادبی به یک مشکل بزرگ برای ناشران، محققان و موسسات آموزشی تبدیل شود. اگر سرقت ادبی به درستی شناسایی نشود، متقلبان و سارقان می‌توانند به نتایجی برسند که مستحق آن نیستند. انواع مختلف سرقت ادبی وجود دارد. کپی-جایگزینی، بازگردانی، سرقت ادبی ترجمه‌ای، سرقت ادبی هنرمندانه، سرقت ادبی ایده، سرقت ادبی کد و سرقت ادبی اشتباه مرجع دادن[3] از انواع سرقت‌های ادبی به شمار می‌آیند[۲].

امروزه سیستم‌های زیادی مانند [۳] COPS، [۴] SCAM ، [۵] MOSS و ...جهت شناسایی سرقت ادبی ایجاد شده‌اند. بیشتر این سیستم‌ها براساس ساختار لغوی و الگوریتم‌های تطابق رشته عمل می‌کنند. بنابراین این سیستم‌ها به سختی می‌توانند سرقت‌های بازگردانی و جایگذاری مترادف‌ها را شناسایی کنند.شناسایی سرقت‌های ادبی معنایی چالش بزرگی است که در بحث شناسایی سرقت ادبی وجود دارد. در این مقاله برای بهبود چالش موجود در مبحث شناسایی سرقت ادبی از برچسب‌گذاری نقش‌معنایی به همراه الگوریتم ژنتیک استفاده شده است.

برچسب‌گذاری نقش‌معنایی[4] جهت تعیین نقش هر کلمه در جمله استفاده می‌شود. در این مقاله از فرهنگ لغت WordNet استفاده شده است.

با استفاده از فرهنگ لغت WordNet مجموعه مترادف‌های هر کلمه در جمله استخراج می‌شود. این روش قادر است که سرقت‌های کپی-جایگزینی، بازگردانی یا جایگذاری مترادف‌ها، تغییر ساختار کلمات در جمله را شناسایی‌کند. قسمت‌های بعدی مقاله به صورت زیر سازماندهی می‌شود. در قسمت دوم توصیف کاملی از کارهای مرتبط با شناسایی سرقت ادبی ارائه می‌شود. در قسمت سوم برچسب‌گذاری معنایی شرح داده می‌شود. درقسمت چهارم الگوریتم ژنتیک شرح داده می‌شود. در قسمت پنجم معیار شناسایی شباهت معرفی می‌شود. در قسمت ششم روش

1Text mining
2Plagiarism Detection
3Dangling
4Semantic Role Labeling

پیشنهادی که مبتنی بر برچسب‌گذاری نقش‌معنایی و الگوریتم ژنتیک است بیان می‌شود. طراحی آزمایش‌ها و مجموعه داده‌های PAN-PC-09 و نتایج در قسمت هفتم بحث می‌شود. قسمت هشتم نتیجه‌گیری مقاله را بیان می‌کند.

کارهای مرتبط

بسیاری از روش‌های شناسایی سرقت ادبی به ویژگی‌های لغوی مبتنی بر کاراکتر، ویژگی‌های لغوی مبتنی بر کلمه و ویژگی‌های نحوی وابسته‌اند. تطابق رشته بین دو رشته به این معنا است که آن‌ها کاراکترهای دقیقا یکسان دارند. برخی روش‌ها تطابق رشته‌ها و شناسایی سرقت ادبی را برمبنای درصد اثر انگشت پیدا می‌کنند[۶ و ۷ و ۸]. این روش‌ها کارایی خوبی دارند اما هنگامی‌که متنی به سرقت می‌رود و قسمت‌هایی از آن بازگردانی می‌شود یا برخی از کلمات متن با مترادف‌های کلمه جایگزین می‌شود، شناسایی سرقت ادبی با روش اثر انگشت با شکست مواجه می‌شود.

روش‌های شناسایی سرقت ادبی معمول برای مقایسه‌ی متن مشکوک به سرقت و متن اصلی مبتنی بر تطابق رشته‌ای است. رشته‌های مشابه با استفاده از تطابق رشته تا حدودی می‌توانند شناسایی شوند و مقایسه‌ی متن‌ها را براساس n-gram در نظر می‌گیرد.

متن مشکوک به سرقت با دو مجموعه از Tri-gram ترکیب می‌شود تا اعمال مقایسه انجام گیرد. Tri-gramها جهت شناسایی سرقت ادبی به کار گرفته می‌شوند. برای اینکه شباهت‌ها به صورت محلی مقایسه شوند، جملات را به عنوان واحدهای مقایسه در نظر می‌گیرد. البته این مقدار مقایسه‌ی بین جملات دقیق کپی شده، درج کلمات دیگر، حذف کردن کلمات دیگر و آوردن جمله با همان معنی اما ساختارمتفاوت، متفاوت خواهد بود[۶].

امروزه بسیاری از سیستم‌های شناسایی سرقت براساس ویژگی‌های نحوی عمل می‌کنند. در این روش با استفاده از برچسب گذاری قسمتی از متن و برخی از معیارهای شناسایی شباهت رشته‌ای به محاسبه‌ی شباهت بین دو متن پرداخته می‌شود. دو متن شبیه به هم از لحاظ نحوی ساختار نحوی برخی از جمله‌هایشان شبیه به هم است. برخی از روش‌ها اسناد را براساس برچسب گذاری قسمتی از متن رتبه بندی می‌کند[۷].

روش‌هایی شناسایی سرقت ادبی که تاکنون معرفی شدند، مبتنی بر کاراکتر بودند. بسیاری از مطالعات بر روی ویژگی‌های ساختاری متن مانند سرفصل، بخش، پاراگراف و مراجع تمرکز کرده‌اند. درخت ساختاریافته‌ی ویژگی‌ها به عنوان یکی از مطالعات اخیر ارائه شده‌است که بر مبنای ویژگی ساختاری است. یک مدل چند لایه‌ای خود سازمان دهنده[5] برای اسنادمطرح می‌شود. این ایده بر اساس دو مرحله ساخته شد. این دو مرحله شامل لایه‌ی بالایی و لایه‌ی پایینی است. لایه‌ی بالایی خوشه بندی و بازگشت اسناد را ارائه می‌دهد. این کار هنگامی که لایه‌ی پایینی مشغول محاسبه‌ی ضریب شباهت کسینوس برای به دست آوردن شباهت و سرقت متن است، انجام می‌شود[۸].

ویژگی‌های لغوی و نحوی جملات می‌توانند توسط برداری از واژه‌ها نشان داده شوند. شباهت بین بردارهاتوسط بردار ضریب شباهت محاسبه می‌شود. در واقع جملات توسط بردارها نمایش داده می‌شوند. بنابراین شباهت بین بردارها می‌توانند توسط تطابق، ضریب همپوشانی و ضریب کسینوسی به دست آیند. بنابراین شباهت بین کلمات می‌تواند توسط تطابق، ضریب جاکارد،ضریب دایک[6]، ضریب همپوشانی[7]، ضریب کسینوسی حساب شود. بنابراین استفاده از ضریبی مانند ضریب کسینوسی در کنار معیارهای دیگر می‌تواند یک سیستم شناسایی سرقت موثر را فراهم کند که از آن در مکان‌هایی که امنیت مهم است (مانند ارسال مقاله به کنفرانس‌ها) استفاده شود[۹].

گیپ[8]وهمکارانش یک روش شناسایی سرقت ادبی مبتنی بر نقل قول ایجاد کردند[۱۰]. این روش برای شناسایی اسناد دانشگاهی که بدون ذکر کردن علامت نقل قول ایجاد می‌شوند به کارگرفته می‌شوند. گیپ و همکارانش روشی ایجاد کردند که برای شناسایی سرقت ادبی بر روی الگوهای مشابه در دنباله‌های نقل قول اسناد دانشگاهی تمرکز می‌کند[۱۱].

گرازیا[9] و همکارانش روش مبتنی بر دو زبان برای شناسایی اسناد مشکوک به سرقت از یک زبان اصلی را پیشنهاد دادند. شباهت بین اسناد اصلی و مشکوک به سرقت به وسیله‌ی یک مدل آماری ارزیابی می‌شود. این مدل احتمال اینکه متن مشکوک به سرقت با متن اصلی ارتباطی داشته باشد و احتمال اینکه واژه‌ای در هر دو متن باشد را تخمین می‌زند. روش آن‌ها با یک فرهنگ لغت اسپانیایی و انگلیسی که ترکیب شده‌اند به شناسایی سرقت ادبی می‌پردازد. این رویکرد به محدودیت‌های نوشته‌های دو زبانی نیاز دارد [۱۲].

در روش فازی، طیفی از درجه‌ی شباهت پیاده سازی می‌شود. این طیف مقداری بین صفر تا یک دارد. اگر دو جمله دقیقا مانند هم باشند امتیاز شباهت آن‌ها یک می‌شود و اگر دو جمله دقیقا مخالف هم باشند امتیاز شباهت آن‌ها صفر می‌شود. به هر کلمه از متن یک مجموعه‌ی فازی نسبت داده می‌شود که شامل کلمه و مجموعه مترادف‌های آن است. بنابراین درجه‌ای از شباهت بین لغت‌های موجود در متن و مجموعه‌ی فازی وجود دارد[۹].

روش معنایی بر روی شناسایی شباهت معنایی دو متن کار می‌کند. دو جمله می‌توانند از لحاظ ساختاری و نحوی با هم فرق داشته باشند در حالیکه از لحاظ معنایی شبیه به هم باشند. عثمان و

5ML-SOM

6Dice
7Overlapping
8Gipp
9Grozea

همکارانش یک روش معنایی شناسایی سرقت ادبی مبتنی بر برچسب گذاری معنایی ارائه دادند[۱۳]. بسیاری از تحقیقات علمی جهت شناسایی شباهت معنایی از شبکه واژگان استفاده کرده اند. گلبوخ[۱۰] شباهت معنایی بین کلمات را با محاسبه‌ی میزان ارتباط بین کلمات با استفاده از شبکه واژگان به‌دستآورد[۱۴].

چاو[۱۱] و همکارانش روشی ارائه داده‌اند که شباهت بین متن مشکوک به سرقت و متن اصلی را براساس جملات خبری محاسبه می‌کند[۱۵]. جملات خبری با استفاده از درخت تجزیه استخراج می‌شود. درجه‌ی شباهت بین جملات استخراج شده با استفاده ازفرهنگ لغت محاسبه می‌شود.

الزهرانی[۱۲] و همکارانش روشی ارائه دادند که شباهت معنایی را با استفاده از شباهت رشته‌ای و منطق فازی محاسبه می‌کرد. وقتی دو جمله از متن کاملا شبیه به هم بودند امتیاز یک و در غیر اینصورت امتیاز صفر می‌گرفتند[۱۶]. سرقت‌های معنایی با استفاده از بازگردانی، جایگذاری مترادف‌ها و تبدیل جمله‌ی معلوم به مجهول و بالعکس و ... انجام می‌شود.

سیستم COPS، برمبنای درهم سازی واحدهای جمله است. یک متن ممکن است شامل تعداد زیادی از واحدها باشد.روشی که واحدهای متن را برای محاسبه انتخاب می‌کند، استراتژی انتخاب واحد[۱۳] گفته می‌شود. این سیستم شامل دو تابع است. تابع اول واحدها را از اسناد با توجه به استراتژی انتخاب واحد انتخاب می‌کند و تابع دوم درهم سازی این واحدها را در یک جدول درهم ساز ذخیره می‌کند[۹].

سیستم SCAM، یک سیستم شناسایی سرقت ادبی است که درموسسه‌ی استنفورد توسعه یافته است. بر خلاف COPS این سیستم بر اساس بردار فضای حالت[۱۴] برای ثبت کردن اسناد استفاده می‌کند. این سیستم از یک معیار شباهت جدید که دقت را افزایش می‌دهد استفاده می‌کند. این معیار شباهت بر روی سیستم‌های بازگردانی اطلاعا ت[۱۵] برای جستجوی شباهت معنایی[۱۶] استفاده می‌کند[۹].

سیستم SID، این سیستم در دانشگاه سانتا باربارا[۱۷] توسعه یافته است. طراحان این سیستم شباهت دنباله‌ای را در نظر گرفتند. معیاری که مقدار شباهت بین دو دنباله را اندازه گیری می‌کند بر مبنای پیچیدگی Kolmogorov است. ضمانت جهانی برای این سیستم وجود دارد. یعنی اینکه این سیستم قادر است معیار شباهت را ایجاد کند و بر اساس آن به محاسبه‌ی شباهت بپردازد[۱۷].

سیستم TURNITN، یکی از سرویس‌های رایج شناسایی سرقت ادبی است. این سیستم به دست گروهی به نام iPlagiarism جهت استفاده‌ی معلمان و موسسات آموزشی توسعه داده شد. این سیستم یک سیستم غیر رایگان است و جهت استفاده از این سرویس باید در سایت این سرویس ثبت نام کرد. استادان و معلمان تکالیف دانش آموزان و دانشجویان را به سایت سرویس می‌دهند و نتیجه را یک تا دو روز بعد دریافت می‌کنند. این سیستم تکالیف را با اسناد موجود در وب و پایگاه داده‌های متنی بزرگ مقایسه می‌کند و نتیجه را گزارش می‌کند[۱۷].

ابرتر[۱۸] و همکارانش روشی جهت شناسایی سرقت ادبی با استفاده از سبک بیان نویسنده ارائه داده‌است[۱۸].

استاماتوس[۱۹] و همکارانش یک روش برای شناسایی سرقت ادبی داخلی ارائه داده‌است[۱۹]. سی وارد[۲۰] و همکارانش پیچیدگی Kolmogorov را معرفی کردند که اطلاعات ساختاری متن را جهت شناسایی سرقت ادبی داخلی استخراج می‌کرد[۲۰].

کاسپرزا[۲۱] و همکارانش مدلی برای شناسایی سرقت ادبی خارجی پیشنهاد دادند[۲۱]. این مدل شامل دو فاز است که فاز اول شاخص گذاری روی متن را انجام می‌دهد و فاز دوم به محاسبه ی شباهت بین دو متن می‌پردازد.

زینی[۲۲] و همکارانش برای شناسایی سرقت ادبی از خوشه بندی متن استفاده کرد[۲۲]. او برای شناسایی شباهت بین خوشه‌ها از کلمات کلیدی استفاده کرد. عثمان وهمکارانش از روش‌های مبتنی بر گراف برای شناسایی سرقت ادبی استفاده کردند. به این صورت که مفاهیم دو متن را به شکل گراف پیاده سازی کردند و گراف‌ها را برای شناسایی شباهت اسناد باهم تطبیق دادند و مقایسه کردند[۲۳].

روش ارائه شده در این مقاله براساس روش معنایی به شناسایی سرقت ادبی می‌پردازد. این روش می‌تواند اکثر سرقت‌های ادبی معنایی را تشخیص دهد. روش پیشنهادی با استفاده از برچسب گذاری نقش‌معنایی و فرهنگ لغت WordNet به تجزیه و تحلیل معنایی می‌پردازد.

برچسب‌گذاری نقش‌معنایی

برچسب‌زنی نقش‌معنایی، وظیفه‌ی استخراج نقش‌های معنایی جملات نظیر فاعل، مفعولم ستقیم، مفعول غیرمستقیم، فعل و... را برعهده دارد.

قابها یا ساختارهای معنایی توسط فیلمور[۲۳] معرفی شد[۲۴]. برچسب گذاری نقش‌معنایی کلمات، عملی اساسی برای بسیاری از

10Gelbukh
11Chow
12Alzahrani
13Selecting Chunck
14Space Vector Model
15Information Retrieval
16Semantic Similarity
17Santa Barbara

18Oberreuter
19Stamatatos
20Seaward
21Kasprza
22zini
23Fillmore

حوزه‌های پردازش زبان طبیعی[۲۴] از قبیل ترجمه‌ی ماشینی، شباهت معنایی[۲۵] و ... است. برچسب‌گذاری نقش معنایی در بسیاری از کاربردهای پردازش زبان طبیعی مانندخلاصه سازی [۲۴]، خوشه بندی متن[۲۶]و طبقه بندی متن[۲۷] به کار می‌رود. زمانی که قاب‌های معنایی ایجاد شدند FrameNet نیز توسط باکر[۲۶] کشف شد[۲۸].

سوردن [۲۷] و همکارانش[۲۹]، پرادها ن[۲۸] وهمکارانش[۳۰] و ایکسو[۲۹]وهمکارانش [۲۸] با استفاده‌ی از ماشینهای یادگیرکارایی برچسب گذاری نقش معنایی در کاربردهای متن کاوی بهبود دادند.

کاربردهای وسیعی از برچسب گذاری نقش معنایی مبتنی بر شبکه عصبی برای استخراج اطلاعات از متن‌های پزشکی انجام شد[۳۲]. برچسب گذاری نقش معنایی همچنین برای تحلیل معنایی در شناسایی سرقت ادبی به کاررفته است[۳۳]. برچسب زنی نقش معنایی را تحلیل در سطح جمله در نظر می‌گیرند که درآن فعل جمله مشخص کننده‌ی رویداد واقع شده و سایر اجزای جمله هر یک نقشی در ارتباط با این رویداد می‌پذیرند[۳۴]. برچسب گذاری نقش معنایی با استفاده از شبکه ی عصبی در دو مرحله ای پیاده سازی شده است[۳۵]. عثمان با روشی مرسوم به CHAID در راه بهبود شناسایی سرقت ادبی با استفاده از برچسب گذاری نقش معنایی قدم برداشته است[۳۶]. ابزار زیادی جهت برچسب گذاری نقش معنایی ایجاد شده است. این ابزار شامل [۳۲] VerbNet، FrameNetو Propbank[۲۸]هستند.

در روش پیشنهادشده از ابزار برخط برچسب‌گذاری نقش معنایی دانشگاه Illinois جهت برچسب‌گذاری نقش‌معنایی استفاده شد.

الگوریتم ژنتیک

الگوریتم ژنتیک[۳۰]روش جستجویی در علم رایانه برای یافتن راه‌حل تقریبی برای بهینه‌سازی و مسائل جستجو است. الگوریتم‌های ژنتیک هنگام جستجو در فضای حالت، نتایج خوبی به همراه دارند [۳۳]. با توجه به اینکه الگوریتم ژنتیک برای بهینه سازی به کارمی‌رود استفاده از آن در حوزه‌ی بازیابی اطلاعات مانند شاخص گذاری اتوماتیک اسناد[۳۹]، طبقه بندی اسناد[۴۰] و خلاصه سازی متن [۴۱]و ... در سال‌های اخیر روند رو به رشدی داشته است. بوآرارا[۳۱] و همکارانش روشی جهت شناسایی سرقت ادبی مبتنی بر الگوریتم ژنتیک در سروهای ایمیل ارائه دادند[۴۲].

شبکه‌های عصبی مصنوعی و الگوریتم ژنتیک جهت شناسایی سرقت ادبی براساس سبک نویسنده استفاده شدند[۴۳].

عامل اصلی انتقال خصوصیات زیست‌شناختی در موجودات زنده کروموزوم[۳۲] و ژن[۳۳]است. نحوه عملکرد الگوریتم ژنتیک به گونه‌ای است که در نهایت ژن‌ها و کروموزوم‌های برتر مانده و ژن‌های ضعیف‌تر از بین می‌روند. به عبارت دیگرنتیجه‌ی عملیات متقابل ژن‌ها و کروموزوم‌ها باقی‌ماندن موجودات اَصلح و برتر می‌باشد[۴۴].

قبل از اینکه الگوریتم ژنتیک مورد استفاده قرارگیرد باید پارامترهای مسئله‌ای که قرار است بهینه شود، کدگذاری شود. هر کروموزوم شامل ژن‌هایی است که در کنار هم جمع شده‌اند. هر ژن یک نقشدر جمله را نشان می‌دهد. در این روش دو نوع کروموزوم Original و Suspect جمعیت اولیه را تشکیل می‌دهند. روش کار در این مقاله به این صورت است که یک متن اصلی را با ۹ متن مشکوک به سرقت مورد مقایسه قرار می‌دهد و تعداد جملات به سرقت رفته را شناسایی می‌کند. جمعیت اولیه به این صورت تشکیل می‌شود که متن اصلی هر بار با یکی از متون مشکوک به سرقت مقایسه می‌شود. بعد از عمل پیش پردازش یک جمله از متن اصلی با تمام جملات متن مشکوک به سرقت مورد مقایسه قرار می‌گیرد. این عمل به صورت تکراری انجام می‌شود تا تمام جملات متن اصلی با تمام جملات متن مشکوک به سرقت مقایسه شود. بنابراین جمعیت اولیه در هر بار تکرار(به تعداد جملات متن اصلی و متن مشکوک به سرقت) شامل ۱۶کروموزوم (۱کروموزوم ازمتن اصلی، ۷ کروموزوم مترادف اصلی، ۱ کروموزوم از متن مشکوک به سرقت و ۷ کروموزوم مشتق شده ی مشکوک) است.

تابع ارزیابی توسط تابع محاسبه ی شباهت WU- PALMER پیاده سازی می‌شود. که در ادامه راجع به آن بحث شده است.

روش انتخاب کروموزوم به این صورت است که کروموزوم‌هایی که بیشترین امتیاز شباهت را کسب می‌کنند انتخاب می‌شوند. روش انتخاب از نوع نخبه گرایی است. در این روش بهترین فرزندان برای تولید نسل بعد نگهداری می‌شوند.

تقاطع[۳۴]یکی از عملگرهایالگوریتم ژنتیک است. در روش پیشنهادی از تقاطع یکنواخت استفاده می‌شود. در این روش یک رشته به طول رشته والد انتخاب می‌شود. اگر عدد محاسبه ی شباهت از Pm کمتر باشد این مقدار Φ و اگر بزرگتر یا مساوی pm باشد مقدارمترادف قرار می گیرد.

تابع محاسبه‌ی شباهت

در روش پیشنهادی برای محاسبه‌ی شباهت بین مجموعه مترادف‌ها و نقش از جمله‌ی مشکوک به سرقت از تابع هدف یا

24Natural Language Processing
25Semantic Similarity
26Baker
27Surdeanu
28pradhan
29Xue
30Genetic Algorithm
31Bouarara

32Chromosome
33Gen
34Crossover

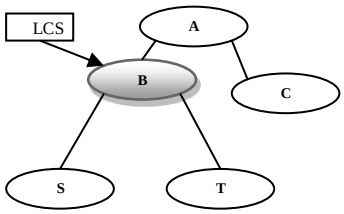
تابع محاسبه‌ی شباهت WU- PALMER استفاده می‌شود که در ادامه هر تابع با جزییات بیشتری بیان می‌شود.

معیار محاسبه‌ی شباهت WU-PALMER برمبنای طول مسیراست. شمارش تعداد گره‌های درخت یا پیوند بین گره‌ها، یکی از راه‌های ممکن جهت محاسبه‌ی شباهت معنایی است. کمترین فاصله بین دو مفهوم به معنای بیشترین شباهت بین آن‌ها است[۱۴].

معیار WU-PALMER براساس طول عمق و نزدیک‌ترین رده بند مشترک مجموعه مترادف‌ها، به محاسبه‌ی شباهت می‌پردازد. رابطه‌ی۱نحوه‌ی محاسبه شباهت را براساسWU-PALMER نشان می‌دهد.

$$sim(S,T) = (2 * Depth(LCS))/(Depth(S) + Depth(T)) \quad (1)$$

در رابطه‌ی۱،T را به مجموعه مترادف‌های کلمه از متن اصلی و Sبه مجموعه مترادف‌های کلمه از متن مشکوک به سرقت اشاره می‌کند. Depth (T) به کوتاهترین فاصله از گره ریشه تا گره اشاره می‌کند. Depth (S) به کوتاهترین فاصله از گره ریشه تا گره اشاره می‌کند. LCS[۳۵] به نزدیک‌ترین رده بند مشترک بین T و S اشاره می‌کند. باید توجه داشت که LCS به معنای طولانی‌ترین زیر دنباله‌ی مشترک نمی‌باشد. در واقع نزدیکترین رده بند مشترک به معنای جد دو مفهوم از دو مجموعه مترادف می‌باشد. مثلا LCS {truck, car} = automotive است. شکل(۱) درخت مفاهیم را در شبکه واژگان WordNet نشان می‌دهد.



شکل ۱. درخت مفاهیم درشبکه واژگان**WordNet**

شناسایی سرقت ادبی مبتنی بر برچسب گذاری نقش معنایی والگوریتم ژنتیک

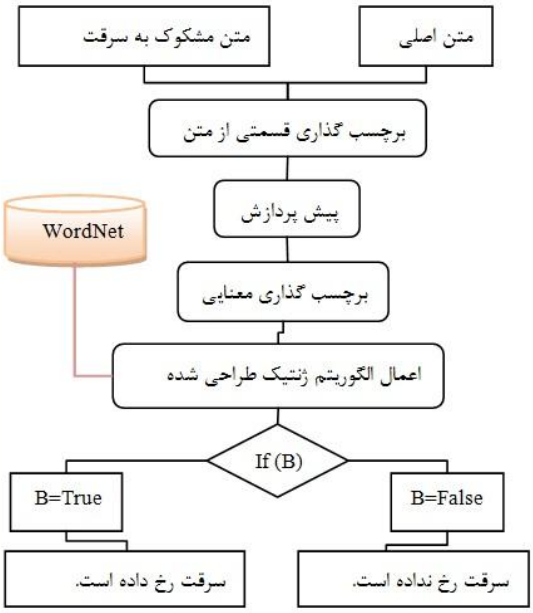
در این قسمت بر روی ایده‌ی روش پیشنهادی صحبت می‌شود. در ابتدا برچسب گذاری قسمتی از متن انجام می‌شود. گام دوم روش پیشنهادی پیش‌پردازش است. سپس عمل پیش پردازش که شامل قطعه‌کردن متن، حذف ایست‌واژه‌ها و ریشه‌گیری است، انجام می‌گیرد. بعد از انجام این مراحل برچسب گذاری نقش معنایی برای کلمات جملات متن اصلی و متن مشکوک به سرقت انجام می‌شود.ARG0، ARG1، ARG2، V، NEG، ADJ و DIR برچسب‌هایی است که به کلمات جملات خورده می‌شود. جدول(۱) نشان می‌دهد که هر کدام از این برچسب‌ها در جمله چه نقشی را نشان می‌دهند. قرار گرفتن حرف O و S برای مشخص نمودن برچسب‌های جمله ی اصلی و مشکوک به سرقت است.گام بعدی اعمال الگوریتم ژنتیک است. الگوریتم ژنتیک خود شامل دو مرحله‌ی استخراج مفاهیم و گروه‌بندی نقش‌ها و محاسبه‌ی تابع شباهت و ارزیابی است.

جدول۱.برچسب کلمات در متن به همراه نقششان در جمله

توضیح	برچسب
فاعل	**ARG0**
مفعول مستقیم/ضمیر مفعولی	**ARG1**
نقش‌های اضافی	**ARG2**
فعل	**V**
علامات منفی در جملات	**NEG**
صفت	**ADJ**
حروف اضافه	**DIR**

بعد از برچسب‌گذاری معنایی، الگوریتم ژنتیک بر روی داده‌های غیرساخت‌یافته یا همان لغات متن انجام می‌گیرد. در روش پیشنهادی جمله‌ها، کروموزوم‌ها و نقش‌های جملات ژن‌ها هستند. دو کروموزوم با هفت ژن در نظر گرفته می‌شود. یک کروموزوم برای نقش‌های جمله‌ی اصلی و یک کروموزوم برای نقش‌های جمله‌ی مشکوک به سرقت در نظر گرفته می‌شود. ژن‌ها همان نقش‌های کلمات در جمله‌ی اصلی و جمله‌ی مشکوک به سرقت است. سپس همه‌ی مفاهیم و مترادف‌ها را برای هر واژه‌ی برچسب خورده در جمله‌ی اصلی را با استفاده از فرهنگ لغت WordNet استخراج می‌شود. این مفاهیم برای هر نقش از جمله استخراج می‌شوند. سپس الگوریتم ژنتیک طراحی شده را بر روی ژن‌های جمله‌ی مشکوک به سرقت و مجموعه مترادف‌های ژن‌های جمله‌ی اصلی که در کروموزم قرار می‌گیرد را اعمال می‌شود. شکل (۲) معماری کلی روش پیشنهادشده را نشان می‌دهد. در ادامه معماری روش پیشنهادی با جزئیات بیشتری توضیح داده می‌شود.

35least common sub-sumer



شکل ۲.معماری روش پیشنهادی

برچسب گذاری قسمتی از متن

بخش‌های سخن، طبقه‌بندی‌هایی زبانی از کلمات هستند که رفتار نحوی یک قسمت از جمله را بیان می‌کنند. مهمترین بخش‌های نحوی اسم، فعل، صفت و قید است. روش پیشنهادی از ابزار برخط دانشگاه Illinois برای برچسب گذاری قسمتی از متن[۳۶] استفاده می‌کند.

پیش پردازش

پیش پردازش یکی از گام‌های مهم در پردازش زبان طبیعی و متن کاوی است. پیش پردازش شامل مراحل قطعه‌کردن، حذف ایست واژه‌ها و ریشه‌گیری است. برای متن کاوی نیاز است که بر روی متن عملیاتی انجام شود. یکی از گام‌های اولیه پیش پردازش، قطعه کردن متن است.در این گام متن به واحدهای معنی دار مانند جملات،کلمات و... تقسیم می‌شود. در روش پیشنهادشده واحدهای معنی‌دارجمله‌ها هستند. جملات همان قطعه‌های معنی‌دار هستند. این گام برای متن اصلی و متن مشکوک به سرقت انجام می‌شود.

ایست‌واژه‌ها[۳۷] کلماتی هستند که در متن زیاد تکرار می‌شوند. این کلمات معنای مفیدی ندارند و سرعت پردازش را کاهش می‌دهند. فضای خالی زیادی از حذف کردن ایست‌واژه‌ها ایجاد می‌شود اما حذف این کلمات تاثیری در بازگردانی اطلاعات ندارد. در متن‌کاوی، ایست‌واژه‌ها حذف می‌شوند. حذف این کلمات نتایج پردازش را بهبود می‌دهد و سبب کاهش بار محاسبات و افزایش سرعت خواهد شد. به همین دلیل،این کلمات غالبا در فاز پیش پردازش حذف می‌شوند.

بعد از حذف ایست‌واژه‌ها گام بعدی ریشه‌یابی است. در این مرحله به منظور یکسان سازی اشکال مختلف یک کلمه، یکپارچه سازی انجام می‌شود. ریشه‌یابی به فرایند تبدیل کلمات به فرم ریشه‌ای و پایه‌ای آن‌ها اشاره می‌کند. این مرحله در پردازش متن اهمیت زیادی دارد زیرا باعث می‌شود کامپیوتر با کلمات هم‌خانواده که ظاهرا با هم متفاوت هستند مانند دو کلمه‌ای که که از لحاظ ریشه‌ای هیچ ارتباطی با هم ندارند، برخورد ننماید. الگوریتم‌های مختلفی برای ریشه‌یابی لغات پیشنهاد شده است. روش پیشنهادی از الگوریتم پورتر برای ریشه‌یابی لغات متن استفاده می‌کند.

برچسب‌گذاری نقش‌معنایی جمله‌های پیش پردازش شده

در این مرحله برچسب گذاری نقش معنایی برای جملات متن اصلی و متن مشکوک به سرقت انجام می‌شود. با استفاده از برچسب‌گذاری نقش معنایی نقش هر کلمه به آن برچسب می‌خورد. برچسب‌گذاری کلمات را با استفاده از سرویس برخط دانشگاه Illinois انجام شده است. شکل (۳) نحوه برچسب گذاری را با استفاده از این سرویس برای جمله‌ی The book was given to Mary by John نشان می‌دهد.

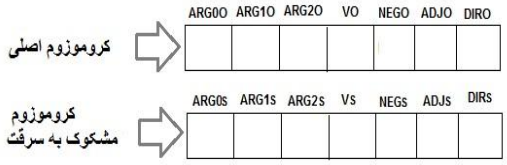
Output:

	SRL	Charniak
The	thing given [A1]	(S1 (S (NP (DT The)
book		(NN book))
was		(VP (AUX was)
given	V: give	(VP (VBN given)
to	entity given to [A2]	(PP (TO to)
Mary		(NP (NNP Mary)))
by	giver [A0]	(PP (IN by)
John		(NP (NNP John)))))

شکل ۳. برچسب گذاری معنایی

استخراج مفاهیم و گروه‌بندی نقش‌ها

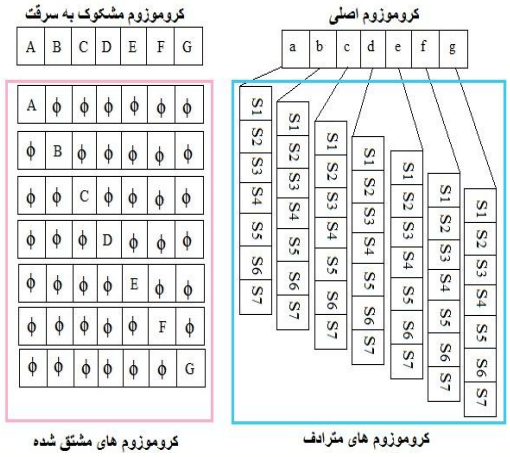
در این مرحله نقش‌های جملات متن اصلی و متن مشکوک به سرقت گروه بندی می‌شود. برای گروه‌بندی دو کروموزوم Suspect, Original با اندازه‌ی هفت یا هفت ژن در نظر گرفته می‌شود. هر کدام از ژن‌ها مربوط به یک نقش در جمله می‌شوند. در شکل(۴) این اختصاص نشان داده شده است.



شکل ۴.اختصاص نقش‌های جملات به ژن‌های کروموزوم

36Part of Speech Tagging
37Stop Word Removal

بعد از اینکه جایگاه هر ژن درکروموزوم‌ها مشخص شد با استفاده از فرهنگ لغت WordNet به استخراج مفاهیم می‌پردازیم. در روش پیشنهادی، کروموزوم‌ها همان جمله‌ها هستند. نحوه‌ی مقایسه متن مشکوک به سرقت و متن اصلی به صورت جمله به جمله انجام می‌گیرد. برای هر ژن از کروموزوم جمله‌ی اصلی، مفاهیم و مجموعه مترادف‌ها را با استفاده از WordNet به دست می‌آید. مجموعه مترادف‌ها هر نقش از جمله‌ی اصلی، خود یک کروموزوم جداگانه تشکیل می‌دهند. این کروموزوم‌ها، کروموزوم‌های مترادف[38] نامگذاری شده است. کروموزوم حاصل از جمله‌ی مشکوک به سرقت نیز هفت ژن دارد. با توجه به تعداد نقش‌های تعریف شده برای روش پیشنهادی که هفت نقش است، هفت کروموزوم دیگر (کروموزوم‌های مشتق‌شده) تعریف می‌شود. این کروموزوم‌ها فقط یک ژن دارند که با رشته‌یϕ جایگزین نشده است. ژنی که با رشته‌یϕ جایگزین نشده است همان نقش از جمله‌ی مشکوک به سرقت یا ژن از جمله‌ی مشکوک به سرقت است که با توجه به جایگاه نقش در کروموزوم اولیه جمله‌ی مشکوک به سرقت ژن غیرتهی است. هر بار در یک جایگاه قرار می‌گیرد. شکل(۵) نحوه نمایش کروموزوم‌ها برای جمله‌ی متن اصلی و مشکوک به سرقت و کروموزوم‌های مترادف را نمایش می‌دهد.



شکل ۵.نمایش کروموزوم‌ها برای جمله‌ی متن اصلی و مشکوک به سرقت و کروموزوم‌های مترادف

دو آرایه از نوع عدد صحیح با اندازه‌ی ۷×۱ با نام‌های Scorex و Scorex' در نظر گرفته می‌شود. مقادیر حاصل از امتیازات مقایسه‌ی شباهت دو ژن در این آرایه‌ها ذخیره می‌شود.یک متغیر از نوع رشته‌ای با نام Suspect نیز در نظر گرفته می‌شود که در طول انجام الگوریتم این متغیر هر بار یکی از ژن‌های غیرتهی کروموزوم مشکوک به سرقت را برای مقایسه با ژن‌های مجموعه مترادف کروموزوم اصلی در خود نگهداری می‌کند.

در این روش تابع هدف همان تابع محاسبه‌ی شباهت WUPALMER است. در واقع با به کارگیری این تابع به محاسبه‌ی امتیاز شباهت بین نقش از جمله‌ی مشکوک و مجموعه مترادف‌ها از جمله‌ی اصلی پرداخته می‌شود.

هفت متغیر بولی با نام‌های B_1، B_2، B_3، B_4، B_5، B_6 و B_7 تعریف می‌شود. هر کدام از این متغیرهابیولی مربوط به یک نقش (ژن) از جمله‌ی مشکوک به سرقت (کروموزوم مشکوک به سرقت) است.

محاسبه‌ی شباهت با استفاده از توابع شباهت(تابع هدف) و تعریف تابع ارزیابی

تابع ارزیابی با تغییراتی بر روی تابع هدف انجام تعریف می‌شود. اولین تابع ارزیابی در رابطه۲ نشان داده شده است.

$$Scorex[\,] = \begin{cases} SIM(S,T) * 0 & Scorex[i] < 0.5 \\ SIM(S,T) * 1 & Scorex[i] \geq 0.5 \end{cases} \quad (2)$$

در رابطه۲، Scorex[i] امتیاز شباهت مقایسه دو ژن در اولین مرحله (ژن‌هایی که درکروموزوم مترادف است با ژنی که مقدارش در متغیر Suspect ذخیره می‌شود) است. بعد از اینکه آرایه Scorex پر شد، در این مرحله تابع ارزیابی به کارگرفته می‌شود. در این مرحله با کمک آرایه‌ی Scorex اگر ژنی از مجموعه مترادف‌ها امتیاز شباهتش با متغیر Suspect کمتر از ۰,۵ باشد، امتیاز شباهتش صفر می‌شود و خود نیز از کروموزوم مترادف حذف می‌شود و به جای آن رشته‌یϕ قرار می‌گیرد. اگر امتیاز محاسبه‌ی شباهت در این مرحله بزرگتر یا مساوی ۰,۵ باشد، امتیازش را در یک ضرب کرده و ژن مذکور در جای خود باقی خواهد ماند.به عبارت دیگر علاوه بر به دست آوردن تابع ارزیابی، جهش نیز صورت خواهد گرفت. علاوه بر این از مکانیسم نخبه گرایی نیز استفاده می‌شود. زیرا ژن‌هایی انتخاب می‌شوند که نسبت به بقیه شباهت بیشتری با ژن درون متغیر Suspect دارند.

در مرحله‌ی بعد دوباره کروموزوم مترادف به روز شده با ژنی که مقدارش در متغیر Suspect ذخیره می‌شود مقایسه می‌شود و امتیاز حاصل از مقایسه‌ی ژن‌های کروموزوم مترادف و ژنی که مقدارش در متغیر Suspect ذخیره می‌شود این بار در آرایه‌ی scorex' ذخیره می‌شود. در این مرحله دومین تابع ارزیابی به صورت رابطه‌ی۳ تعریف می‌شود.

$$Scorex' = \begin{cases} SIM(S,T) * 0 & Scorex' < 1 \\ SIM(S,T) * 1 & Scorex' \geq 1 \end{cases} \quad (3)$$

آنگاه مطابق با آرایه‌ی'Scorex اگر ژنی از مجموعه مترادف‌ها امتیاز شباهتش کمتر از ۱ باشد،امتیاز شباهتش صفر می‌شود و خود نیز از کروموزوم مترادف حذف می‌شود و به جای آن رشته‌یϕ قرار

38SynSet Chromosome

می‌گیرد. اگر امتیاز مقایسه شباهت در این مرحله مساوی ۱ باشد، امتیازش را در یک ضرب کرده و ژن مذکور در جای خود باقی خواهد ماند. آنگاه مطابق با کروموزوم مترادفی که در حین اجرای الگوریتم دوبار به روز می‌شود اگر ژنی در این کروموزوم باشد که مقدار غیرتهی داشته باشد به معنای این است که ژن از کروموزوم مشکوک به سرقت با ژن از کروموزوم اصلی هم معنا است.

سپس از هفت متغیر بولی تعریف شده متغیری که به ژن مشکوک به سرقت (نقش کلمه در جمله) اختصاص داده شده است مقدار یک می‌گیرد و در غیراینصورت مقدار صفر می‌گیرد. این فرایند برای سایر نقش‌ها یا ژن‌ها از کروموزوم مشکوک به سرقت انجام می‌شود.

شکل (۶، ۷ و ۸) نحوه‌ی اجرای الگوریتم ژنتیک روش پیشنهادی را بر روی کروموزوم‌ها به صورت نمونه نشان می‌دهد.

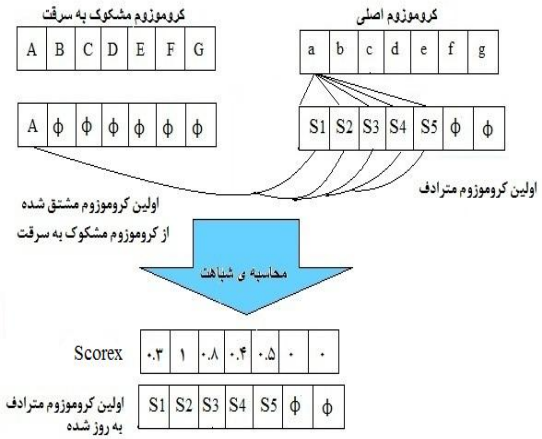
لازم به ذکر است که در شکل (۶،۷و ۸) فرض شده است که دو ژن هم‌معنا باشند. برای اینکه دو کروموزوم مقایسه شوند، کل ژن‌های کروموزوم‌ها باید مورد بررسی قرار گیرند. شرط خاتمه ی الگوریتم ژنتیک یا حد آستانه ی تابع ارزیابی این است که هفت مقدار بولیB1 الی B7 به دست آید.

گاهی اوقات ممکن است که تمام ژن‌های کروموزوم مترادف پر نشوند زیرا ممکن است که تعداد مفاهیم و مجموعه مترادف‌ها کمتر از هفت عدد باشد. لذا به جای ژن‌های خالی کروموزوم مترادف مقدار رشته یϕ قرار می‌گیرد.

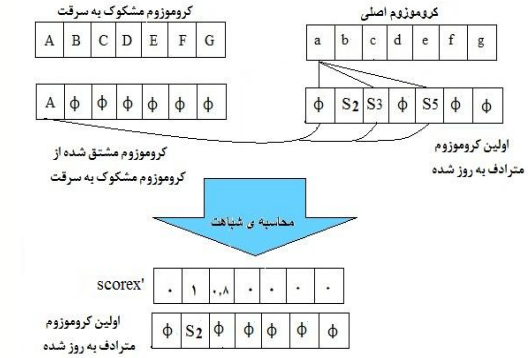
مقادیر همه‌ی متغیرهای بولی بررسی می‌شود. اگر همه‌ی متغیرهای بولی مقدار درست بگیرند به معنای این است که دو کروموزوم (جمله) مورد مقایسه با هم، هم‌معنی بوده‌اند و سرقت تشخیص داده می‌شود، در غیر اینصورت سرقت ادبی رخ نداده است. مقدار B مطابق با رابطه ی ۴ حساب می‌شود.

$$B = B1 + B2 + B3 + B4 + B5 + B6 + B7 \qquad (4)$$

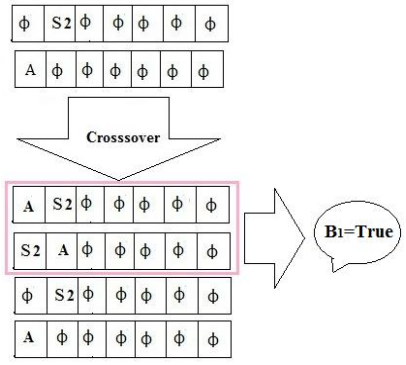
اگر دو کروموزوم یا دو جمله هم‌معنا باشند به تعداد جملات به سرقت رفته یکی افزوده می‌شود.



شکل ۶. اجرای فاز اول الگوریتم ژنتیک



شکل ۷. اجرای فاز دومالگوریتم ژنتیک



شکل ۸.اجرای فاز نهایی الگوریتم

شبه کد روش پیشنهادی در شکل (۹) آورده شده‌است:

```
Input:
population size 16N, crossover probability pc,
termination condition;
Output: plagiarism detect results;
Step 1: Text preprocessing;
Step 2: Semantic role labeling;
Step 3: Text representation;
Step 4: Encoding;
Step 5: Initialize population. Chromosomes of
original and suspect text;
Step 6: Calculate the similarity function;
Step 7: Calculate the fitness value of all individuals;
Step 8: Selection, crossover;
Step 9: repeat from 4;
Step 10: Calculate B;// end of repeat and
termination
condition
Step 11: End algorithm, output the plagiarism
detection.
```

شکل۹. شبه کد روش پیشنهادی

نتایج

در این مقاله از مجموعه داده‌های PAN-PC-9 برای ارزیابی روش پیشنهادی استفاده شده‌است. برای ارزیابی روش پیشنهادی از سه پارامتر Recall، Precision و F-measure استفاده شده‌است. برای بار اول ارزیابی سه متن مشکوک به سرقت را با متن اصلی مقایسه می‌شود. برای بار دوم ارزیابی را با شش متن مشکوک به سرقت و متن اصلی انجام می‌شود. برای آخرین بار نه متن مشکوک به سرقت را با متن اصلی مقایسه می‌شود و نتایج مورد ارزیابی قرار می‌گیرد.

روش پیشنهادی بر روی تعداد متنهای مختلف (۹،۶،۳) آزمایش شد. برای ارزیابی روش پیشنهادی سه پارامتر که معمولا در شناسایی سرقت ادبی کاربرد دارند، در رابطه۵،۶ و۷ بیان شده‌اند.

(۵)

$$\text{Recall} = \frac{\text{NumberofDetectSentence}}{\text{TotalNumberofSentence}}$$

(۶)

$$\text{Precision} = \frac{\text{NumberofPlagiarizedSentence}}{\text{NumberofDetectSentence}}$$

(۷)

$$\text{F} - \text{Measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

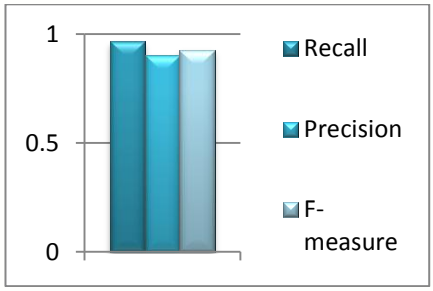
جدول (۲) مقدار پارامترهای ارزیابی را برای ۹متن با استفاده از روش پیشنهادی نشان می‌دهد. در این جدول نتایج مقایسه ی ۹ متن مشکوک به سرقت با یک متن اصلی آورده شده است.

جدول (۳) به ترتیب میانگین پارامترهای ارزیابی را برای شناسایی شباهت سه، شش و نه متن مشکوک به سرقت را با متن اصلی نشان می‌دهد. برای اولین بار سه متن مشکوک به سرقت با یک متن اصلی مقایسه می‌شود و برای هر کدام از سه متن مقدارهای پارامترهای ارزیابی ثبت می‌شود و در نهایت از آن میانگین گرفته می‌شود. برای بار دوم ۶ متن مشکوک به سرقت را با متن اصلی مقایسه شده و میانگین‌پارامترهای ارزیابی گرفته می‌شود. برای سومین بار ۹ متن مشکوک به سرقت را با متن اصلی مقایسه کرده و میانگین پارامترهای ارزیابی گرفته می‌شود.

جدول۳. میانگین پارامترهای ارزیابی

تعداد متن مشکوک به سرقت	Recall	Precision	F-measure
۳	۰,۹۸۴	۰,۹۰۵	۰,۹۳۹
۶	۰,۹۷۳	۰,۹۰۲	۰,۹۳۴
۹	۰,۹۶۶	۰,۹۰۰	۰,۹۲۷

شکل (۱۰) میانگین پارامترهای ارزیابی روش پیشنهادی را برای ۹ متن نشان می‌دهد.



شکل۱۰. میانگین پارامترهای ارزیابی روش پیشنهادی

در جدول (۴) به عنوان نمونه متن سوم و متن هشتم مشکوک به سرقت با متن اصلی مورد مقایسه قرار گرفته است و پارامترهای الگوریتم ژنتیک در آن درج شده است.

جدول ۴. پارامترهای الگوریتم ژنتیک در ارزیابی متن سوم و متن هشتم مشکوک به سرقت با متن اصلی

پارامترها	متن سوم	متن هشتم
اندازه ی جمعیت اولیه	۴۶۴	۸۳۲
نوع تقاطع	یکنواخت	یکنواخت
نرخ تقاطع	۱	۱
اندازه ی پنالتی	۱ یا ۰	۰ یا ۱
ماکزیمم نسل	۷۴۲۴	۱۳۳۱۲
برازندگی در اولین تکرار	۰,۸۱	۰,۶۵
برازندگی در دومین تکرار	۰,۸۹	۰,۷۲

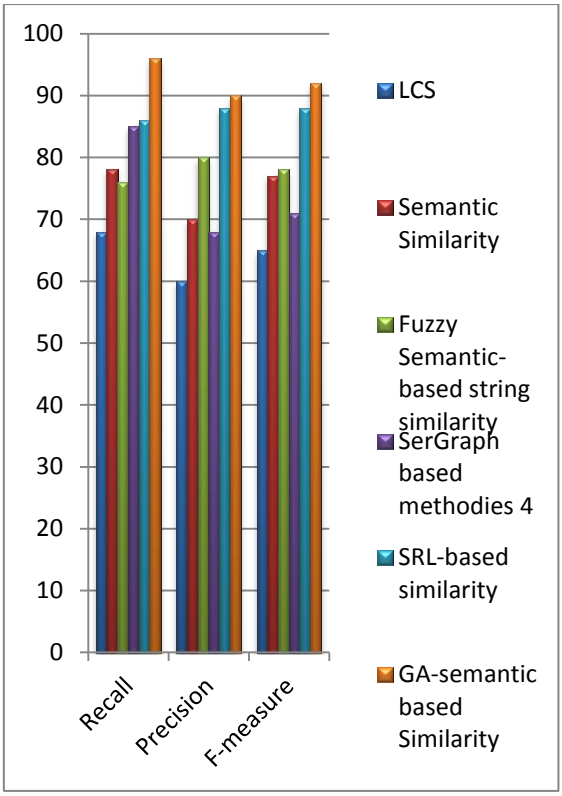
همانطور که در جدول (۴) نشان داده شده است برازندگی در هر بار تکرار اصلاح می‌شود و شباهت بیشتری را نشان می‌دهد. با استفاده از برازندگی در تکرار اول بهترین کروموزوم‌ها برای تکرار دوم در نظر گرفته می‌شوند. در تکرار دوم دیده می‌شود که برازندگی به بیشترین حد خود رسیده است.

شکل (۱۱) مقایسه‌ی بین روش پیشنهاد شده را با روش‌های مبتنی برگرا ف[۳۹] [۲۳]، مبتنی بر برچسب‌گذاری نقش معنایی[۴۰] [۱۳]، مبتنی بر طولانی ترین زیر دنباله ی مشترک[۴۱] [۴۵]، روش شباهت رشته ای مبتنی بر معنایی فازی[۴۲][۱۶]، مبتنی بر روش معنایی[۴۳][۱۵]را شرح می‌دهد.

با توجه به شکل (۱۱) نتیجه‌ای که گرفته می‌شود این است که در روش پیشنهادی پارامترهای ارزیابی شناسایی سرقت ادبی نتایج بهتری را کسب می‌کنند. علاوه براین روش پیشنهادی می‌تواندانواع سرقت ادبی کپی-جایگزینی، جایگذاری مترادف‌ها، تغییر ساختار جملات و سرقت‌های معنایی را شناسایی کند.

با توجه به نتایج به دست آمده میانگین معیارهای ارزیابی بالای ۰,۷۰ است. این مقدار به نظر می‌رسد که نتیجه ی خوبی باشد چرا که بالاتر از ۰,۵۰ است. بنابراین از نظر کارایی نسبت به روش‌های ارائه شده بهتر عمل می‌کند.

روش‌هایی که در شکل (۱۱) با روش پیشنهادی مقایسه شده اند دارای پیچیدگی زمانی $O(n^2)$هستند. روش پیشنهاد شده در این مقاله نیز دارای پیچیدگی زمانی $O(n^2)$ است. البته این پیچدگی زمانی به عنوان یکی از چالش‌هایروش‌های شناسایی سرقت ادبی محسوب می‌شود.



شکل ۱۱. مقایسه ی روش پیشنهادی و روش‌های ارائه شده قبلی

روش پیشنهادی صرفا برای مقالات علمی و متن‌های خبری کاربرد دارد. یکی از مشکلاتی که روش پیشنهاد شده دارد و باعث می‌شود که دقت روش پیشنهادی پایین بیاید این است که در هنگام شناسایی جملات متن، جملاتی که داری کلمات ربطی‌اندبه‌خوبی نمی‌تواند تشخیص دهد و آن‌ها را مورد پردازش قرار نمی‌دهد و آنها را یک جمله در نظر می‌گیرد. به همین علت همانطور که در جدول (۲) نشان داده شده است در شناسایی برخی از جملات با مشکل مواجه است.

نتیجه گیری

که نتایج بهینه‌تر شود. در مقالات آینده سعی خواهد شد که پیچیدگی زمانی این الگوریتم کمتر شود و مشکل مربوط به شناسایی دقیق تعداد جملات متون حل شود.

در این مقاله یک روش جهت بهبود روش‌های شناسایی سرقت ادبی مبتنی بر الگوریتم ژنتیک و برچسب‌گذاری نقش معنایی ارائه شد. در روش پیشنهادی، برچسب گذاری نقش معنایی به همراه الگوریتم ژنتیک می‌تواند کارایی خوبی در شناسایی سرقت ادبی داشته باشد. الگوریتم ژنتیک ساختاری است که با استفاده از آن روش پیشنهادی بهینه می‌شود. استفاده‌ی همزمان از برچسب‌گذاری نقش‌معنایی و برچسب‌گذاری قسمتی ازمتن باعث

39Graph-Based Method
40Semantic Role Labeling -Based
41Longest Common Subsequence
42Fuzzy Semantic-Based String Similarity
43Semantic-Based Similarity

شد که دقت روش پیشنهادی نسبت به روش‌هایی که قبلا ارائه شده‌اند بیشتر باشد.

روش پیشنهادی نسبت به روش مبتنی برگراف، مبتنی بر برچسب گذاری نقش معنایی،مبتنی بر طولانی ترین زیر دنباله ی مشترک، روش شباهت رشته ای مبتنی بر معنایی فازی، مبتنی بر روش معنایی کارایی بهتری دارد. روش پیشنهادی می‌تواند سرقت‌های کپی-جایگزینی، جایگذاری مترادف‌ها، تغییر ساختار جملات و سرقت‌های معنایی و ... را تشخیص دهد.

در ادامه این مقاله و برای کارهای بعدی سعی می‌شود که به هر کدام از نقش‌های جمله وزن داده شود. با وزن‌دهی به کلمات و به کارگیری الگوریتم‌هایی مانند الگوریتم ژنتیک سعی می‌شود

جدول ۵. ارزیابی پارامترهای پیشنهادی با ۹ متن

تعداد اسناد متن مشکوک	تعداد کل جملاتمتن مشکوک	تعدادجملات سرقت رفت	تعدادجملات به سرقت نرفته	تعداد جملات شناسایی شده	تعداد جملات شناسایی نشده متن	Recall	Precision	F-measure
متن اول	۴۵	۴۲	۳	۴۵	۰	**۱**	۰,۹۳۳	۰,۹۶۵
متن دوم	۱۵۷	۱۳۹	۱۸	۱۵۶	۱	۰,۹۸۷	۰,۸۹۱	۰,۹۳۶
متن سوم	۲۹	۲۵	۴	۲۸	۱	۰,۹۶۵	۰,۸۹۲	۰,۹۱۶
متن چهارم	۴۲	۳۶	۱۶	۴۰	۲	۰,۹۵۲	۰,۹۰۰	۰,۹۲۴
متن پنجم	۶۹	۶۰	۲۹	۶۸	۱	۰,۹۸۵	۰,۸۸۲	۰,۹۲۹
متن ششم	۷۴	۶۵	۲۴	۷۲	۲	۰,۹۷۲	۰,۹۰۲	۰,۹۳۵
متن هفتم	۴۹	۴۶	۳	۴۹	۰	۱	۰,۹۳۸	۰,۹۶۸
متن هشتم	۵۲	۳۵	۱۷	۴۸	۴	۰,۹۲۳	۰,۷۲۹	۰,۸۱۴
متن نهم	۴۳	۴۰	۱۷	۴۲	۱	۰,۹۷۶	۰,۹۵۲	۰,۹۶۳

مرجع‌ها

[1] A. S.BIN-HABTOOR,M.A.ZAHER. (2012). A Survey on Plagiarism Detection Systems. *International Journal ofComputer Theory and Engineering, 4* (2), 185-188.

[2] A. Si, H.V. Leong, R.W.H. Lau. (1997). CHECK: a document plagiarism detection system. *Proceedings of the 1997 ACM symposium on Applied computing* (pp. 70-77). San jose,CA,United states: ACM.

[3] A.H. OSMAN, N. SALIM , M.S. BINWAHLAN,R.Alteeb,A.Abuobieda. (2012). An improved plagiarism detection scheme based on semantic role labeling. *Applied Soft Computing, 12*, ۱۴۹۳-۱۵۰۲.

[4] A.H. Osman, N. Salim. (2013). An Improved Semantic Plagiarism Detection Scheme Based on Chi-squared Automatic Interaction Detection. *INTERNATIONAL CONFERENCE ON COMPUTING, ELECTRICAL AND ELECTRONIC ENGINEERING (ICCEEE)* (pp. 640-647). Khartoum : IEEE.

[5] A.H. Osman, N. Salim, M.S. Binwahlan, S. Twaha,Y.J. Kumar,A. Abuobieda. (2012). Plagiarism Detection Scheme Based on Semantic RoleLabeling. *The International Conference on Information Retrieval and Knowledge Management, CAMP' ۱۲ (Capaian Maklumat dan Pengurusan Pengetahuan)* (pp. 30-33). Kuala Lumpur: IEEE.

[6] A.H. Osman, N. Salim,M .S. BinWahlan, H. Hentabli, A. Ali. (2011). Conceptual similarity and graph-based method for plagiarism detection. *Journal of Theoretical and Applied Information Technology, 32* (2), 135-145.

[7] A.Z, B. (1997). On the resemblance and containment of documents. *Compression and Complexity of Sequences Proceedings* (pp. 21-29). Salerno : IEEE .

[8] B. GIPP, J. BEEL. (2010). Citiation based plagiarism detection:a new approach to identify plagiarized work language independently. *Proceedings of the 21th ACM Conference on Hyptetext and hypermedia* (pp. 273-274). New York, NY, USA: ACM.

[9] B. GIPP, N. MEUSCHKE. (2011). Citation pattern matching algorithms for citation-based plagiarism detection. *Proceedings of the 11th ACM symposium on Document engineering* (pp. 249-258). New York, NY, USA : ACM.

[10] C. Grozea,C.Geh, M. Popescu. (2009). ENCOPLOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection. *Workshop "Uncovering Plagiarism, Authorship and Social Software Misuse"* (pp. 10-18). Donostia, Spain, 2009: ENCOPLOT.

[11] C.F. Baker, C.J. Fillmore, J.B. Lowe. (1998). The Berkeley FrameNet Project. *Proceedings of the 17th International Conference on Computational Linguistics.1*, pp. 86-90. Montreal, Quebec, Canada: Association for Computational Linguistics.

[12] C.J.Fillmore. (1968). The case for case. In C.J.Fillmore. New York: Holt, Rinehart, and Winston: In Bach and Harms (Ed.): Universals in Linguistic Theory.

[13] C.K.kent, N. (2010). Features based text similarity detection. *Journal of Computing, 2* (1), 53-57.

[14] C.K.kent, N. salim. (2010). Web based cross language plagiarism detection. *Second International Conference on Computational Intelligence, Modelling and Simulation.* (pp. 199-204). Bali : IEEE.

[15] C.Lyon, ,J.A.Malcolm,R.G.Dickerson. (2001). Detecting Short Passages of Similar text in large document. *the 2001 Conference on Empirical Methods in Natural Language Processing* (pp. 1-8). New York: Cornell University.

[16] D.WHITLEY. (1994). A Genetic Algorithm Tutorial. *Statics and Computing, 4* (2), 65-85.

[17] E. Stamatatos,B. Stein, P. Rosso, M. Koppel, E. Agirre. (2009). Intrinsic plagiarism detection using character n-gram profiles. *SEPLN ۲۰۰۹workshop on uncovering plagiarism, authorship, and social software misuse (PAN ۰۹)*, (pp. 38-46). San Sebastian.

[18] E.R.Fonseca,J.L.G.Rosa. (2013). A Two-Step Convolutional Neural Network Approach for Semantic Role Labeling. *The 2013 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-7). Dallas, TX: IEEE.

[19] G.Oberreuter, J.D. Vel squez. (2013). Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style. *Expert Systems with Applications, 40* (9), 3756-3763.

[20] H.A. Bouarara,R.M. Hamou. (2014). Machine Learning Tool and Meta-heuristic Based On Genetic Algorithms For Plagiarism Detection Over Mail Service. *2014 IEEE/ACIS 13th International Conference on Computer and Information Science (ICIS)* (pp. 157-162). Taiyuan, China: IEEE.

[21] H.Uguz. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems, 24* (7), 1024-1032.

[22] J.Dierderich. (2006). Computational Methods to Detect Plagiarism in Assessment. *7th International Conference on Information Technology Based Higher Education and Training (ITHET '۰۶)* (pp. 147-154). Ultimo, NSW : IEEE.

[23] J.Kasprzak,M. Brandejs. (2010). Improving the reliability of the plagiarism detection system. *Notebook Papers of CLEF 2010 LABs and Workshops.* Padua, Italy: ACM.

[24] K. Kipper, H.T. Dang, M. Palmer. (2000). Class-based construction of a verb lexicon. *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence* (pp. 691–696). Menlo Park, California: AAAI Press.

[25] K.Monostor, A. (2000). Document overlap detection system for distributed digital libraries.

Proceedings of fifth ACM conference on Digital libraries (pp. 226-227). San Antonio, TX, United States: ACM.

[26] L. Suanmali, N. Salim, M.S. Binwahlan. (2009). Automatic text summarization using feature-based fuzzy extraction . *Jurnal Teknologi Maklumat, 2* (1), 105-155.

[27] L.s, B. T. (2011). A Study of Plagiarism Detection Tools and Technologies. *International Journal of Advanced Research in Technologies, 1* (1), 64-70.

[28] L.Seaward,S. Matwin. (2009). Intrinsic plagiarism detection using complexity analysis. *SEPLN workshop on uncovering plagiarism, authorship, and social software misuse (PAN ۰۹)*, (pp. 56-61). San Sebastian.

[29] M. ELHADI, A. AL-TOBI. (2008). Use of text syntactical structures in detection of document duplicates. *Digital Information Management Third International Conference on ICDIM* (pp. 520-525). London: IEEE .

[30] M. Surdeanu,S.Harabagiu,J.Williams,P.Aarseth. (2003). Using predicate-argument structures for information extraction. *Proceedings of the 41st Annual Meeting on Association forComputationalLinguistics* (pp. 8-15). Sapporo, Japan: Association for Computational Linguistics.

[31] M. Zini, M. M. (2006). Plagiarism detection through multilevel text comparison. *Automated Production of Cross Media Content for Multi-Channel Distribution Second International Conference on AXMEDIS '06* (pp. 181-185). Leeds: IEEE.

[32] M.A. Fattah,F.Ren. (2008). Automatic Text Summarization. *World Academy of Science, Engineering and Technology, 27*, 192-195.

[33] N. Shivakumar, H. Garcia-Molina. (1995). SCAM: a copy detection mechanism for digital documents. *2nd International Conference in Theory and Practice of Digital libararies* (pp. 1-12). Austin, TX: Stanford University.

[34] N. Xue, M. Palmer. (2004). Calibrating features for semantic role labeling. *Conference on Empirical Methods in Natural Language Processing* , (pp. 88-94). University of Pennsylvania, Philadelphia PA.

[35] N.Heintze. (1996). Scalable document fingerprinting. *UNIX Workshop on Electronic Commerce*, (pp. 191-200). Oakland, California.

[36] S. ALZAHRANI, N. SALIM. (2010). Fuzzy Semantic-based Sring Similarity for Extrinsic Plagiarism Detection. *Braschler and Harman* , 1-8.

[37] S. Brin, J. Davis, H. Garcia-Molina. (1995). Copy detection mechanisms for digital documents. *SIGMOD '95 Proceedings of the 1995 ACM SIGMOD international conference on Management of data* (pp. 398-409). San Jose, CA, United States: ACM.

[38] S. Pradhan, W. Ward, K. Hacioglu, J.H. Martin, D. Jurafsky. (2004). Shallow semantic parsing using support vector machines. *Proceedings of NAACL-HLT* (pp. 233-240). Boston, Massachusetts, USA: Association for Computational Linguistics.

[39] S. Shehata, F. Karray, M.S. Kamel. (2010). An efficient model for enhancing text categorization using sentence semantics. *Computational Intelligence, 26* (3), 215-231.

[40] S.M.Krishna, S. M. (2010). An efficient approach for text clustering based on frequent itemsets. *European Journal of Scientific Research, 42* (3), 399-410.

[41] S.R.Rastgar,N.Razavi. (2013). *A System for Building Corpus Annotated With Semantic Roles.* MASTER THESIS, Swedish.

[42] S.Torres, A. (2009). Computing Similarity Measures for Original WSD Lesk Algorithm. *Advances in Computer Science and Application, 43*, 155-166.

[43] S.Y. Yuen,C.K. Chow. (2009). A Genetic Algorithm that adaptively mutates and never revisits. *Transactions on Evolutionary Computation, 13* (2), 454-472.

[44] T. BARNICKEL, J. WESTON, R. COLLOBERT, H. MEWES, V.STUMPFLEN. (2009). Large scal application of neural network based semantic role labeling for automated relation extraction from biomedical texts. *International Journal of Information Technology & Decisionn Making, 4* (7), 6393.

[45] T.W.S.Chow,M.K.M.Rahman. (2009). Multilayer SOM With tree-Structured Efficient Document Retrieval. *Transactions on Neural Networks, 20* (9), 1385-1402.

[46] W. Song, S.C. Park. (2009). Genetic algorithm for text clustering based on latent semantic indexing. *Computers & Mathematics with Applications, 57* (11-12), 1901-1907.

[٤٧] آ.ک. قالیباف، س.ر. قوچانی،ا.استاجی. (۸۸). برچسب زنی نقش معنایی جملات فارسی با رویکر یادگیری مبتنی بر حافظه. فصل نامه ی پردازش علائم و داده ها, ۱ (۱۱), ص. ۱۳-۲۲.

[٤٨] رضوان یعقوبی،حسن ختن لو و منصور اسماعیلپور. (۱۱ بهمن ۹۲). سرقت ادبی و تقلب علمی:مروری بر روش‌ها،سیستم ها و تکنولوژی های شناسایی سرقت ادبی در مقالات علمی. همایش ملی فن آوری محاسبات و اطلاعات: روندها و سرحدهای جدید (ص. ۴۲). ملایر: دانشگاه ملایر.