تشخیص سرقت علمی اسناد فارسی با رویکرد مبتنی بر گراف

مژگان ممتاز[۱]، مصطفی صالحی[۲]، هادی ویسی[۳]

[۱] دانشجوی کارشناسی ارشد دانشگاه تهران، m.momtaz92@ut.ac.ir

[۲] استادیار، عضو هیئت علمی دانشگاه تهران، mostafa_salehi@ut.ac.ir

[۳] استادیار، عضو هیئت علمی دانشگاه تهران، h.veisi@ut.ac.ir

چکیده

در این مقاله روشی نوین برای تشخیص سرقت علمی در اسناد فارسی ارائه شده است. این روش از ساختار گراف و یکی از روش‌های تشابه گراف (روش تکرار در مشابهت گره‌ها) برای یافتن تشابه در دو سند متنی به زبان فارسی استفاده می‌کند. در این روش ابتدا دوتایی‌های سند مشکوک به سرقت علمی را بدست می‌آوریم و با دوتایی‌های اسناد ذخیره شده در پایگاه داده مقایسه می‌کنیم، اگر تعداد دوتایی‌های مشترک در دو سند از حد آستانه مشخص بیشتر باشد این دو سند برای تشخیص وجود یا عدم وجود سرقت علمی به عنوان ورودی تابع تشخیص سرقت علمی انتخاب می‌شوند. در این تابع ابتدا سندها به گراف‌هایی دارای ساختار منظم تبدیل می‌شوند، سپس به کمک روش تشابه وجود همسایگان مشترک در دو گراف، وجود سرقت علمی یا عدم وجود سرقت علمی، به عنوان خروجی این تابع حاصل می‌شود. پس از اجرای روش پیشنهادی روی دو مجموعه داده، معیار F، این روش نسبت به روش مبتنی بر n-گرام نرم‌افزار مشابه یاب سمیم نور، ۲۰ درصد روی مجموعه داده اول (سرقت علمی از نوع معنایی) و ۱۳ درصد روی مجموعه داده اول (شامل انواع سرقت علمی)، بهبود یافته است. همچنین توانایی این روش برای مقابله با داده نویز بیشتر می‌باشد.

واژه‌های کلیدی

سرقت علمی، گراف، تشابه‌یابی، متن فارسی

۱- مقدمه

امروزه حجم زیادی از اطلاعات را گونه‌های مختلف داده‌های متنی مانند کتاب، مقاله و سایر اسناد تشکیل می‌دهند و حجم این داده‌ها روزانه در حال افزایش است. به‌دلیل افزایش رشد داده‌های متنی، نیاز به رویکردهای جدیدی برای تجزیه و تحلیل، کاوش و استخراج دانش از این نوع داده‌ها وجود دارد. دستیابی به ابزارهای پیشرفته خودکار برای استخراج اطلاعات از داده‌های متنی، یکی از مهمترین اهداف متن‌کاوی[۱] می‌باشد. در تحقیقات اخیر، گراف به عنوان رویکردی جدید برای متن‌کاوی معرفی شده است [1].

ساختارهای مبتنی بر گراف می‌توانند اطلاعات مهم مانند ساختار چیدمان کلمات، جفت کلمات پر تکرار و سایر ویژگی‌های متن را استخراج کند. متن‌کاوی کاربردهای زیادی دارد و یکی از کاربردهای مهم آن شناسایی اسناد مشابه می‌باشد [2]. در بسیاری از موارد نیاز داریم که در میان تعداد زیادی از اسناد، سندهای تکراری و یا نزدیک به کپی را تشخیص دهیم. یکی از شاخه‌های مهم شناسایی اسناد مشابه، تشخیص سرقت علمی در اسناد می‌باشد. سرقت علمی، به معنی استفاده از نوشته‌ها و اطلاعات نویسنده دیگر است، بدون اینکه قدردانی مناسبی از آن نویسنده شده باشد و یا به آن منبع ارجاع شده باشد [3].

امروزه شناسایی سرقت علمی به کمک نرم افزارها آسان‌تر شده است. اما انواع مختلفی از سرقت همچنان موضوع پردردسری می‌باشد. زمانی که در سرقت علمی ساختار سند مرجع با جابجایی کلمات تغییر کرده باشد یا از کلمات مترادف استفاده شده باشد، روش‌های ارائه شده برای تشخیص سرقت علمی کارایی قابل قبول را ندارند. بنابراین نیاز به بهبود روش‌های تشخیص سرقت علمی می‌باشد. انواع سرقت علمی به دو دسته اصلی سرقت علمی تک زبانه[۲] و بین زبانی[۳] تقسیم می‌شود. در این مقاله هدف پیدا کردن سرقت علمی تک زبانه می‌باشد که خود شامل ۴ سطح سرقت علمی می‌شود.

۱. نزدیک به کپی: در این کلاس در سند مشابه قسمت‌هایی از متن بدون تغییر از متن اصلی آورده شده است.

---

[1] Text mining

[2] Mono-lingual

[3] Cross-lingual

۲. نسخه‌برداری سبک: در این کلاس سند مشابه از متن اصلی با وارد کردن تغییرات جزیی (استفاده از کلمات مترادف و تغییرات دستوری) استفاده کرده است.

۳. نسخه‌برداری سنگین: در این کلاس سند از نسخه اصلی بازنویسی کاملی را با تغییر ساختار انجام داده است.

۴. بدون سرقت علمی: در این کلاس سند از نسخه اصلی در تعداد کوتاهی کلمه مانند گزاره ۲۰۰ تا ۳۰۰ کلمه‌ای استفاده کرده است [2].

مجموعه داده جمع‌آوری شده توسط پژوهشکده فناوری و اطلاعات[4] منتشر شده در PAN2015، در بخش بازیابی منبع تشخیص سرقت علمی، شامل ۴ سطح بالا می‌شود.

روش پیشنهادی، با ایده گرفتن از رویکرد گراف، هر سند را به تعدادی بخش با طول مشخص و هر بخش را در صورت لزوم برای بررسی دقیق سرقت علمی با روش رخداد همزمان کلمات در پنجره با اندازه ثابت به گراف تبدیل می‌کند، پس از این مرحله به کمک روش مشابهت گره‌ها میزان تشابه دو گراف را بدست می‌آوریم، اگر میزان تشابه از حد آستانه مشخص بیشتر باشد، آن بخش به عنوان سرقت علمی برچسب می‌خورد.

روش پیشنهادی، نسبت به روش استفاده شده در نرم افزار مشابه‌یاب سمیم‌نور، در بخش تشخیص سرقت علمی با نسخه برداری سنگین ۲۰ درصد بهبود داشته است، که در بخش تحلیل نتایج بررسی شده است. اما به طور خلاصه اینگونه می‌توان بیان نمود که، روش‌های موجود در برابر روش‌های جدید سرقت علمی (استفاده از کلمات مترادف، جابجایی کلمات، تغییر ساختار و سایر موارد مشابه) کارایی قابل قبول را ندارند. بنابراین نیاز به ابزار پیشرفته‌تری می‌باشد که بتواند در برابر روش‌های جدید نیز کارآمد باشد. در این مقاله رویکرد نوینی برای تشخیص سرقت علمی ارائه می‌شود که از نظر دقت، و مقابله با داده نویز کارایی قابل قبولی دارد.

۲- پیشینه موضوع

برای تشخیص سرقت علمی تک زبانه، تاکنون روش‌های گوناگونی ارائه شده است، در این بخش، هر کدام از این روش‌ها، به اختصار توضیح داده شده‌اند.

روش‌های مبتنی بر کاراکتر[5]: که معروف‌ترین آنها، روش اثر انگشت[6] می‌باشد. الگوریتم‌های اثر انگشت متن را به عنوان مجموعه‌ای از کاراکترها در نظر گرفته، سپس کاراکترها را در دسته‌های n کاراکتری تقسیم می‌نمایند، معروفترین آنها ۱۶-گرام، ۸-گرام و ۵- گرام می‌باشند. در این روش درجه شباهت بستگی به تعداد کاراکترهای مشابه در رشته ها دارد. این روش نتایج خوبی را در تشخیص سرقت علمی بدست می آورد اما زمانی که سرقت علمی با بازنویسی یا تغییر برخی کلمات صورت می‌گیرد، در تشخیص سرقت علمی کارا نمی‌باشد [5] [4]. دسته دوم این روش‌ها، مانند الگوریتم‌های اثر انگشت می‌باشند، با این تفاوت که به جای کاراکترها، سند را مجموعه ای از کلمات در نظر می‌گیرند و این کلمات در پنجره‌هایی با اندازه ثابت قرار می‌دهند. میزان تشابه این پنجره‌ها، درجه تشابه اسناد را مشخص می‌نماید [7] [6].

روش‌های مبتنی بر ساختار[7]: در دو روش قبل به ویژگی کلمات موجود در اسناد توجه شده است، ولی در روش‌های مبتنی بر ساختار، به عنوان‌ها، پارگراف‌ها، بخش‌ها و منابع توجه شده است. یکی از معروف‌ترین روش‌های مبتنی بر ساختار، روش ساختار درخت[8] می‌باشد، که اخیرا به آن توجه بسیاری شده است. در روش ساختار درخت، مدل دولایه‌ای تعریف می‌شود، که لایه بالا برای بازیابی اسناد و لایه پایین برای تشخیص سرقت علمی بین اسناد بازیابی شده به روش‌های تشخیص شباهت مانند شباهت کسینوسی[9]، در نظر گرفته شده است [8].

روش مبتنی بر خوشه‌بندی[10]: در این روش اسناد براساس کلمات خاص (یا کلمات کلیدی) خوشه‌بندی می‌شوند. در این روش، هدف بازیابی اسناد مشابه و سرعت بخشیدن به فرآیند تشخیص سرقت علمی است [10] [9].

روش مبتنی بر دستور[11]: در این روش براساس قواعد دستوری، پیش پردازش اولیه برای تشخیص شباهت اسناد صورت می‌گیرد. یکی از مهمترین قواعد دستوری برچسب‌زنی پاره گفتار[12] می‌باشد. در پژوهش انجام شده براساس این روش، پس از برچسب‌زنی پاره گفتار، از تکنیک طولانی‌ترین زیردنباله مشترک[13] در دو سند، برای تشخیص سرقت علمی استفاده می‌کنند [11] [12].

روش مبتنی بر شباهت معنایی[14]: روش تشخیص سرقت علمی اسناد ، مبتنی بر شباهت معنایی می‌باشد که از شبکه واژگان برای یافتن شباهت معنایی استفاده می‌کند. معروفترین شبکه واژگان در زبان انگلیسی، شبکه وردنت[15] می‌باشد. به کمک شبکه وردنت، می‌توان به اطلاعات بیشتری در مورد یک کلمه دست یافت. این روش، زمانی که سرقت علمی، به کمک استفاده از کلمات مترادف، صورت گرفته باشد، کارا می‌باشد. شبکه واژگان فارس‌نت[16] نیز برای زبان فارسی، جمع‌آوری شده است [14] [13] [15].

روش مبتنی بر گراف[17]: در این روش، هر متن به یک گراف تبدیل می‌شود، که در این گراف گره‌ها می‌توانند کلمات، عبارت‌های اسمی یا جملات موجود در متن باشند و یال‌ها که نشان‌دهنده ارتباط بین گره‌ها می‌باشند می‌توانند ارتباط معنایی بین کلمات یا رخداد هم‌زمان کلمات در یک جمله را نشان دهند. در بخش روش پیشنهادی این روش بیشتر توضیح داده شده است. با تبدیل هر متن به یک گراف می‌توان از مزیت الگوریتم‌های تشابه گراف، برای تشخیص سرقت علمی استفاده نمود [3].

---

4 ICT Research Institute
5 Character-Based
6 Fingerprint
7 Structural-Based
8 Tree-Structured
9 Cosine similarity
10 Classification
11 Syntax-Based
12 Part-of-speech (POS)
13 Longest Common Subsequence (LCS)
14 Semantic-Based
15 WordNet
16 FarsNet

روش پیشنهاد شده در این مقاله، ترکیبی از روش مبتنی بر کاراکتر و روش مبتنی بر گراف می‌باشد. توجه به ساختار متن در این روش باعث می شود حتی در صورتی که سرقت علمی انجام شده از نوع تغییر ساختار هم باشد، سرقت علمی در سند مورد نظر شناسایی شود.

مشکلات روش‌های پیشین در متن‌کاوی، انگیزه‌ای برای ارائه روش‌های جدید برای نمایش متن بوده است. روش‌های پیشین، مبتنی بر روش پایه، یعنی مدل مجموعه کلمات می‌باشند، در نتیجه به ترتیب کلمات بی‌توجه هستند. با این فرض که ترتیب رخداد کلمه در جمله یا متن تاثیری در معنای آن ندارد، این روش‌ها در کاربرد بازیابی اطلاعات، با فرض اینکه ترتیب رخداد کلمات تاثیری در جمله یا متن و معنای آن نداشته باشند، نتیجه ی ضمنی و خوبی را استخراج می کنند. مشکل این روش‌ها در پیدا کردن شباهت قسمت‌های مختلف متن می‌باشد، اگر موضوعی با کلمات دیگر نوشته شود و از لحاظ معنی با متن‌های قبلی مشابهت داشته باشد ، دیگر این روش‌ها برای تشخیص شباهت مناسب نیستند. این روش‌ها همچنین معنی و ساختار متن را بیان نمی‌کنند [16].

۳- روش پیشنهادی برای تشخیص سرقت علمی

هر متن را می‌توان با یک گراف متناظر کرد. استفاده از گراف برای نمایش متن به این دلیل اهمیت دارد که می‌توان یک متن بدون ساختار را به کمک گراف ساختارمند کرد و از مزایای رویکرد گراف آن برای خلاصه‌سازی متن، تشخیص شباهت اسناد و سایر کاربردهای متن‌کاوی بهره‌مند شد. همچنین برای پردازش زبان طبیعی توسط الگوریتم‌ها، نیاز به وجود گراف متن می‌باشد. در گراف متناظر متن ، گره‌ها شامل کلمات و جملات هستند و یال‌های گراف نمایانگر ارتباط بین کلمات می‌باشند، که این ارتباط از روش‌های متفاوتی که بستگی به کاربرد گراف دارد، استنباط می‌گردد. شکل ۱ نمونه‌ای از گراف متناظر با یک متن کوتاه را نمایش می‌دهد.

در این روش، بین هر دو کلمه که فاصله بین آنها کمتر از اندازه پنجره باشد، یال ایجاد می‌شود. روش پیشنهاد شده، برای تشخیص سرقت علمی ۵ مرحله را شامل می‌شود، که در ادامه این مراحل به تفکیک شرح داده می‌شوند.

مرحله ۱- پیش پردازش: ابتدا متن مشکوک به سرقت علمی را نرمال می‌نماییم.

مرحله ۲- انتخاب سندهای کاندید: این متن نرمال با تمام اسناد نرمال شده مرجع در مجموعه داده مقایسه می‌شود. در این مقایسه از بین تمامی اسناد موجود در مرجع، تعدادی از آنها به عنوان سند کاندید، برای بررسی بیشتر انتخاب می‌شوند. در واقع برای بالا بردن سرعت، اسناد موجود در پایگاه داده را فیلتر می‌کنیم. برای فیلترکردن اسناد و انتخاب اسناد کاندید، از بسته هضم [17]، استفاده نمودیم. این بسته برای پردازش زبان فارسی ارائه شده است، که یکی از قابلیت‌های آن برچسب زنی سند فارسی می‌باشد. به کمک این بسته، کلمات موجود در اسناد را برچسب نحوی می‌زنیم، کلماتی که برچسب آنها اهمیت کمی دارد مانند حروف اضافه، قیدها و کلمات ربط را حذف می‌نماییم. پس‌ازاین مرحله تعداد دوتایی‌ها[18] موجود برای کلمات باقیمانده در سند را محاسبه می‌کنیم. در انتها تمامی دوتایی‌ها موجود در همه اسناد مرجع را در فایل متنی به‌همراه نام فایل آنها ذخیره می‌کنیم. همین روند را برای سند مشکوک به سرقت علمی انجام می‌دهیم. اگر تعداد دوتایی‌ها مشترک در سند مرجع (S2) و سند مشکوک به سرقت(S1) از حد آستانه مشخص شده (Ω) بیشتر باشد، سند مرجع به عنوان یکی از کاندیدها برای بررسی بیشتر انتخاب می‌شود.

مجموعه دوتایی‌ها سند مشکوک به سرقت علمی :S1

مجموعه دوتایی‌ها سند مرجع :S2

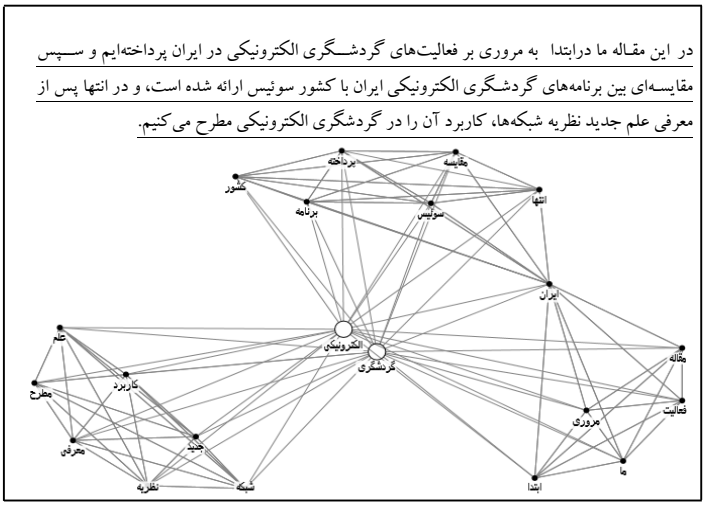
تعداد دوتایی‌های مشترک=$|S1 \cap S2|$

مرحله ۳- تبدیل متن به بخش‌هایی با طول مساوی: سند مشکوک به سرقت و یکی از اسناد کاندید به بخش‌هایی با تعداد کلمات مشخص (۵۰ کلمه در هر بخش) تقسیم می‌شوند.

هر بخش با تمامی بخش‌های سند مرجع مقایسه می‌شود. در این مرحله نیز فیلترگذاری روی بخش‌ها صورت می‌گیرد، تا زمان اجرا کاهش یابد. در این مرحله اگر دو بخش حداقل ۵ کلمه اصلی و یکتا مشترک داشته باشند به مرحله تشکیل گراف می‌روند، درغیراینصورت به بخش بعدی سند مرجع می‌رویم.

مرحله ۴- ساخت گراف متناظر: در مرحله تشکیل گراف هر بخش به گرافی تبدیل می‌شود که گره‌های آن کلمات اصلی و یکتا می‌باشند و در این گراف هر کلمه با ۳ کلمه بعد از خود همسایه می‌شود؛ در واقع بین آن کلمه و سه کلمه بعد از آن به‌ترتیب یال برقرار می‌کنیم.

مرحله ۵- تشخیص سرقت علمی: پس از تشکیل گراف به‌دنبال گره‌هایی در سند مرجع هستیم که با گره ای از گراف سند مرجع مشترک باشد، سپس با استفاده از شباهت جاکارد مشابهت آن را به کمک رابطه (۱) محاسبه می‌کنیم.



شکل ۱- گراف متناظر با یک متن نمونه

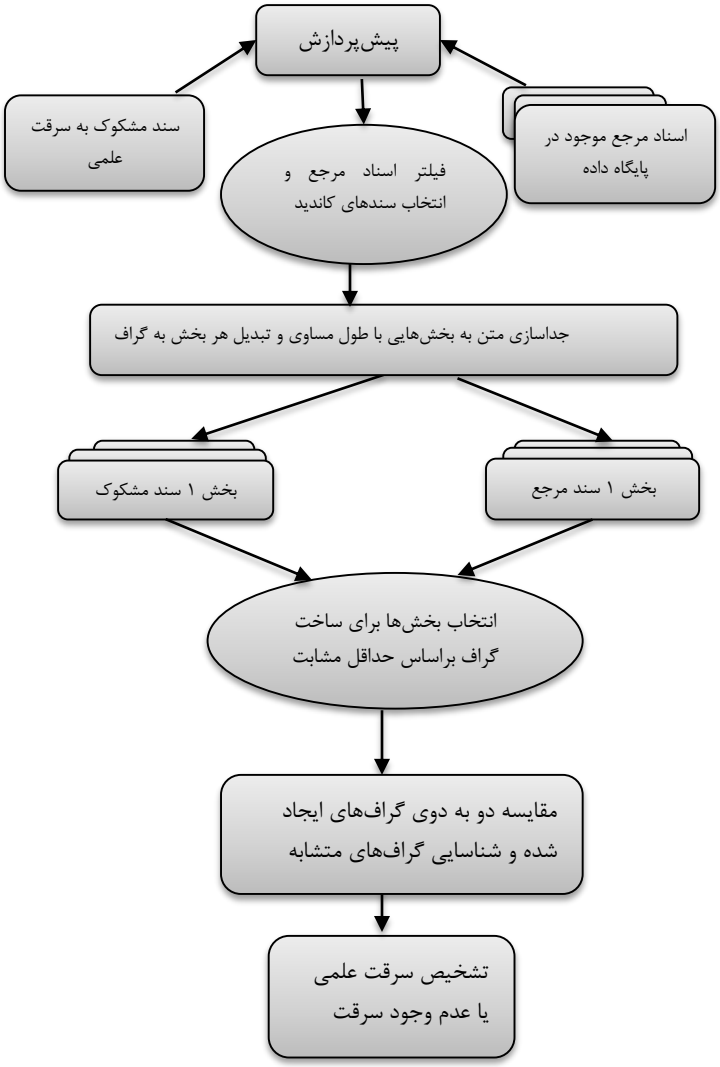
[18] 2-gram

$$similarity(A,B) = \frac{A \cap B}{\max(len(A), len(B))} \quad (1)$$

همسایگان درجه اول گره مشترک گراف سند مرجع :B
همسایگان درجه اول گره مشترک گراف سند مشکوک :A

اگر مشابهت دو گره از حد آستانه α بیشتر باشد (α=0.66)، آن گره به عنوان گره مشابه انتخاب می‌شود. در انتها اگر بخشی بیش از حد آستانه β گره مشابه داشته باشد (β=6)، آن بخش به‌عنوان بخشی که در آن سرقت علمی رخ داده است، برچسب می‌خورد. حد آستانه α و β براساس اجراهای صورت گرفته و به صورت تجربی بدست آمده‌اند.

در شکل ۲ نمودار مراحل تشخیص سرقت علمی که در بالا بیان شدند، آورده شده است.



شکل ۲- نمودار مراحل تشخیص سرقت علمی

۴- پیاده‌سازی و تحلیل نتایج

در این بخش نتایج پیاده‌سازی و اجرای روش پیشنهادی، روی مجموعه داده سرقت علمی اسناد فارسی آورده شده است.

۴-۱. مجموعه داده آزمون

- مجموعه داده ۱: مجموعه داده انتخاب شده شامل ۱۱۲ سند ارائه شده در مسابقه PAN-CELF2015 [18]، است که سرقت علمی آنها از نوع شبیه‌سازی شده و تصادفی با درجه بالا می‌باشد.
- مجموعه داده ۲: مجموعه داده انتخاب شده شامل ۳۰۰ سند مشکوک و ۱۵۰۰ سند مرجع است که در مسابقه PAN-CELF2015 [18]، برای سرقت علمی در زبان فارسی تایید شده است. این مجموعه داده شامل سرقت علمی کپی، سرقت علمی پراکنده تصادفی و سرقت علمی شبیه‌سازی شده است.

۴-۲. ارزیابی کارایی روش پیشنهادی

در این روش پارامترهای مختلفی در نتیجه تاثیر گذار هستند که عبارتند از:

۱- حد آستانه α، β و Ω
۲- روش ساخت گراف (جهت‌دار یا غیرجهت‌دار بودن همچنین تعداد همسایه در نظر گرفته شده برای هر گره)

در تشخیص سرقت علمی در مرحله بازیابی سند، حد آستانه Ω بسیار بیشتر از سایر حد آستانه‌ها در نتیجه بدست آمده موثر بوده است. در جدول۱ نتایج بدست آمده براساس Ωهای متفاوت روی مجموعه داده ۱ آورده شده است. می‌توان مشاهده کرد که در سرقت‌های علمی از نوع معنایی و با درجه ابهام بالا (ابهام به معنای میزان تغییر در ساختار متن اصلی برای انجام سرقت علمی می‌باشد) تعداد دوتایی‌های مشترک، کم می‌باشد، در نتیجه اگر روش‌های مبتنی بر n-گرام به تنهایی مورد استفاده قرار گیرند، بسیاری از این سندها به عنوان سندکاندید انتخاب نخواهند شد.

جدول ۱- کارایی روش پیشنهادی براساس Ωهای متفاوت در مجموعه داده ۱

حدآستانه Ω	دقت	فراخوانی	معیار F
Ω =۲۰	۸۹,۷۰	۴۸,۷۵	۶۳,۱۷
Ω =۱۵	۸۸,۹۵	۵۵,۶۷	۶۸,۴۸
Ω =۱۰	۸۴,۹۱	۶۶,۹۶	۷۴,۸۷
Ω =۶	۷۵,۶	۸۳,۹	۷۹,۵

در جدول ۲ نتایج پیاده‌سازی روش ارائه شده در این مقاله، با روش مبتنی بر n-گرام (روش استفاده شده در نرم افزار تشخیص سرقت علمی سمیم نور) بر روی مجموعه داده ۱ و ۲ آورده شده است.

جدول ۲- مقایسه کارایی روش پیشنهادی با روش مبتنی بر n-گرام

روش	مجموعه داده ۱			مجموعه داده ۲		
	دقت	فراخوانی	معیار F	دقت	فراخوانی	معیار F
روش پیشنهادی	۷۵،۶	۸۳،۹	۷۹،۵	۹۱،۹	۸۳،۹	۸۷،۸
روش n-گرام	۵۵،۶	۵۷،۵	۵۶،۵	۸۴،۷	۶۵،۹	۷۴،۱

۴-۳. تحلیل نتایج

با توجه به نتایج ارائه شده، مشاهده می‌شود که روش پیشنهادی با رویکرد مبتنی بر گراف، نسبت به روش مبتنی بر n-گرام از دقت بالاتری برخوردار است. افزایش دقت به این دلیل است که در روش پیشنهادی متن به بخش‌هایی با تعداد کلمات مشخص تقسیم می‌شود و این امکان را فراهم می‌کند که سرقت‌های علمی جزیی در حدود چند جمله نیز تشخیص داده شوند. همچنین در این روش به جای مقایسه کل متن، فقط بخش‌هایی به دقت بررسی می‌شود که احتمال رخدادن سرقت علمی در آنها زیاد باشد.

به طور کلی، در رویکرد مبتنی بر گراف، بدلیل انعطاف‌پذیری آن در برابر تغییر ساختار، نسبت به روش مبتنی بر n-گرام، برای روش‌های نوین سرقت علمی (که عموما از طریق تغییر ساختار متن، جابجایی کلمات، اضافه‌کردن داده‌های نویز و تغییراتی مشابه با این موارد صورت می‌پذیرند) مناسب‌تر می‌باشد.

۴-۴. ارزیابی روش پیشنهادی با روش‌های مبتنی بر شباهت معنایی

استفاده از شبکه واژگان برای تشخیص سرقت علمی معنایی مناسب است. در زبان انگلیسی شبکه وردنت در ابتدا برای پیدا کردن درجه مشابهت دو کلمه استفاده می‌شود و براساس آن فرمول‌هایی برای تشخیص جملات مشابه ارائه شده است. در این پژوهش هدف ما تشخیص تقلب در زبان فارسی می‌باشد و نسخه اولیه شبکه فارس‌نت بسیاری از امکانات شبکه وردنت را ندارد، همچنین تعداد کلمات موجود در فارس‌نت محدود می‌باشد و این شبکه در حال توسعه می‌باشد، به همین دلیل رویکرد ما در این روش توجه به ساختار متن می‌باشد و امکان مقایسه این روش در زبان فارسی با روش‌های شباهت معنایی امکان پذیر نمی‌باشد [19, 20]. در این روش امکان استفاده از شبکه واژگان برای بهبود دقت تشخیص سرقت علمی وجود دارد و در نظر داریم با بهبود نسخه اولیه فارس‌نت، دقت این روش را برای تشخیص سرقت علمی معنایی بهبود دهیم.

۴-۵. پیچیدگی زمانی روش پیشنهادی

در روش پیشنهادی در مرحله فیلترسازی اولیه، تعداد محدودی از اسناد به عنوان سندهای کاندید انتخاب می‌شوند و سرعت اجرا افزایش می‌یابد. در مرحله بررسی بیشتر تشخیص سرقت علمی(ورودی این قسمت جفت سندها هستند)، به دلیل اینکه سند به بخش‌هایی با طول مساوی تقسیم می‌شوند و برای تبدیل به گراف از فیلتر دیگری گذر می‌کنند(حداقل مشابهت اولیه که براساس کلمات مشترک در دو بخش تعیین می‌شود)، که این عمل نیز تعداد مقایسه کاهش می‌دهد. برای افزایش سرعت زمان اجرا می‌توان هر سند در پایگاه داده را با تعدادی گراف متناظر نمود که هر گراف نماینده یک بخش از سند می‌باشد و در زمان جستجو هر گراف از سند مشکوک مستقیما با یک گراف از پایگاه داده مقایسه شود. در نظر داریم در کارهای آتی با استفاده از تکنولوژی‌های جستجو، زمان اجرا این روش را که در حال حاضر نسبت به روش‌های مبتنی بر n-گرام بیشتر است را بهبود دهیم.

۵- نتیجه‌گیری و کارهای آتی

روش مبتنی بر گراف با تبدیل متن بدون ساختار به متن با ساختار، امکان استفاده از مزایای الگوریتم‌های گراف برای پردازش زبان طبیعی را فراهم می‌نماید. ما این روش را با روش مبتنی بر n-گرام که برای فیلتر اولیه اسناد استفاده می‌شود، ترکیب نمودیم. نتایج حاصل از پیاده‌سازی نشان می‌دهد که این دو رویکرد با هم نتیجه بهتری را ارائه می‌دهند.

در ادامه در نظر داریم که به کمک شبکه واژگان فارس‌نت، دقت الگوریتم را در تشخیص سرقت علمی معنایی افزایش دهیم. همچنین این روش را می‌توان برای ترازبندی بخش‌های سرقت شده تعمیم داد. دسته مهمتری از روش‌های نوین سرقت علمی، سرقت علمی از نوع خلاصه سازی متن اصلی می‌باشد. با توجه به نتایج بدست آمده، پیش‌بینی می‌شود که رویکرد گراف در تشخیص سرقت علمی نوع خلاصه نیز، کارا باشد.

۶- مراجع

[1] Zu Eissen, Sven Meyer, and Benno Stein, "Intrinsic plagiarism detection.", Advances in Information Retrieval, Springer Berlin Heidelberg, . 565-569, 2006.

[2] Oberreuter, Gabriel, and Juan D. VeláSquez, 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, 2013.

[3] N.Kumar, A graph based automatic plagiarism detection technique to handle artificialword reordering and paraphrasing, Computational Linguistics and Intelligent Text Processing,p. 481–494, 2014.

[4] Zini, Manuel, Fabbri, Marco, Moneglia, Massimo, and Panunzi, Alessandro, Plagiarism detection through multilevel text comparison, Second International Conference, IEEE, 2006.

[5] Nahnsen, Thade, Uzuner, Ozlem, and Katz, Boris, Lexical chains and sliding locality windows in content-based text similarity detection, 2005.

[6] Hoad, Timothy C and Zobel, Justin, "Methods for identifying versioned and plagiarized documents," *American society for information science and technolology* , vol. vol. 54(3), p. p. 2013–2015, 2003.

[13] e. a. C. Leacock, "Using corpus statistics and WordNet relations for sense identification," *Comput. Linguist,* vol. vol. 24 , pp. pp. 147-165, 1998.

[14] S. T. a. A. Gelbukh, "Comparing Similarity Measures for Original WSD Lesk Algorithm," *Advances in Computer Science and Application,6,* Vols. vol. 43, pp. 155-16, 2009.

[15] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *Artificial Intelligence Research,* Vols. vol. 11, pp. 95-130, 1999.

[16] Sonawane, S. S., & Kulkarni, P. A, "Graph based Representation and Analysis of Text Document: A Survey of Techniques.," in *International Journal of Computer Applications, 96(19(*, 2014.

[17] "http://www.sobhe.ir/hazm," [Online].

[18] K. e. a. Khoshnavataher, "Developing Monolingual Persian Corpus for Extrinsic Plagiarism Detection Using Artificial Obfuscation," PAN-CELF, 2015.

[19] طاهری ،زهرا و همکاران" تشخیص تقلب متون با استفاده از درخت وابستگی معنایی بین مفاهیم و الگوریتم DTW"، دانشگاه صنعتی شریف،۱۳۹۳.

[20] طاهری ،زهرا و همکاران" تشخیص تقلب متون با استفاده از روابط معنایی بین مفاهیم و الگوریتمDTW "، دانشگاه صنعتی شریف، تهران، ۱۳۹۳.

[7] ". P. t. P. D. ,. C. R. Z. Ceska, "Automatic plagiarism detection based on latent semantic analysis," dissertation, Faculty Appl Sci., Univ. West Bohemia, Pilsen, 2009.

[8] Leilei, K., Haoliang, Q., Shuai, W., Cuixia, D., Suhong, W., & Yong, H, "Approaches for candidate document retrieval and detailed comparison of plagiarism detection," Notebook for PAN at CLEF 2012, 2012.

[9] M. Zini, "Plagiarism Detection through Multilevel Text Comparison," in Automated Production of Cross Media Content for Multi-Channel Distribution," in *Second International Conference on, pp. 181-185*, 2006.

[10] e. a. A. Si, ""CHECK: a document plagiarism detection system," in *presented at the Proceedings of the 1997 ACM symposium on Applied computing*, San Jose, California, United States, 1997.

[11] M. Elhadi and A. Al-Tobi, "Use of text syntactical structures in detection of document duplicates," in Digital Information Management," in *ICDIM 2008. Third International Conference on, , pp. 520-525*, 2008.

[12] M. Elhadi and A. Al-Tobi, "Duplicate Detection in Documents and WebPages Using Improved Longest Common Subsequence and Documents Syntactical Structures," in *presented at the Proceedings of the 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology*, 2009.