کشف دستبرد علمی مبتنی بر شیوه‌های بازیابی اطلاعات

فاطمه مشهدی رجب[1]، مهرنوش شمس‌فرد[2]

[1] دانشکده علوم و مهندسی کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران،
fa.mashhadi@gmail.com

[2] دانشکده علوم و مهندسی کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران
m-shams@sbu.ac.ir

چکیده

اقتباس از آثار علمی دیگران بدون ارجاع صحیح به آنها را دستبرد علمی می‌نامند که کشف خودکار انواع این سوء استفاده‌ها همواره مورد توجه محققین بوده است. در این مقاله روشی جهت کشف دستبرد علمی ارائه شده است که یک روش مبتنی بر بازیابی اطلاعات است. در این روش ما از یک شیوه بازیابی اطلاعات مبتنی بر خوشه‌بندی استفاده کرده‌ایم و در آزمایشات نشان دادیم در سیستم‌های کشف دستبرد علمی، استفاده از شیوه‌های بازیابی اطلاعات مبتنی بر خوشه‌بندی می‌تواند بسیار کاربردی‌تر از شیوه‌های دیگر بازیابی اطلاعات باشد. همچنین در این الگوریتم معیاری برای رتبه‌بندی اسناد بازیابی شده، ارائه شده است. نتایج آزمایشات نشان می‌دهد با استفاده از این معیار، سند مورد نظر در 91٪ موارد در فهرست اسناد رتبه‌بندی شده با رتبه کمتر از پنج حضور دارد. روش کشف دستبرد علمی پیشنهادی قادر به کشف انواع کپی‌برداری‌های دقیق و کپی‌برداری با تغییرات مانند جابجایی جملات، حذف و درج جملات، جایگزینی کلمات با مترادف‌هایشان و ترکیب بخش‌های کپی شده با یکدیگر است. این سیستم قابل توسعه به انواع کپی‌برداری‌های هوشمندانه نیز می‌باشد. در روش پیشنهادی علاوه بر متن اسناد، تصاویر موجود در آنها نیز در رتبه‌بندی اسناد مؤثر خواهند بود. نتایج ارزیابی سیستم پیشنهادی، نشان می‌دهد که در این سیستم برای کپی‌برداری‌های تحت الفظی، میانگین رتبه سند منبع، پنج می‌باشد.

کلمات کلیدی

کشف دستبرد علمی، مشابهت‌یابی متن، خوشه‌بندی اسناد، بازیابی اسناد.

1- مقدمه

در دهه اخیر دستبردهای علمی[1]، در سراسر جهان رشد بسیاری داشته و این امر جوامع علمی را با نگرانی‌های متعددی مواجه ساخته است. در کشور ما نیز مقاله‌ها و پایان‌نامه‌های تقلبی یکی از مشکلات جوامع دانشگاهی به شمار می‌روند. اگر چه عوامل بسیاری در بروز و شیوع این پدیده وجود دارد اما شاید یکی از راهکارهای مقابله با رشد آن، توسعه سیستم‌های کشف دستبردهای علمی در زبان فارسی باشد که می‌تواند از سرعت رشد این پدیده بکاهد.

به طور کلی سیستم‌های کشف دستبرد علمی متون را می‌توان به سه دسته تقسیم کرد[2]:

- سیستم‌هایی که متن سند مشکوک را با مجموعه‌ای از اسناد منبع مقایسه می‌کنند. در این دسته از سیستم‌ها، در دسترس بودن مجموعه‌ای از اسناد منبع لازم و ضروری است[12]، [13]، [14]، [22].
- دسته‌ای دیگر از سیستم‌های کشف دستبرد علمی، سیستم‌هایی هستند که بخش‌های مختلف از سند مشکوک را با یکدیگر مقایسه می‌کنند. در این سیستم‌ها اگر بخشی از سند مشکوک از نظر سبک نوشتاری با بقیه بخش‌ها متفاوت باشد، آن بخش به عنوان یک بخش تقلبی تشخیص داده می‌شود[15] ، [16]، [17]، [21].
- دسته سوم سیستم‌های بین زبانی هستند. در این سیستم‌ها سند مشکوک به زبان مبدأ و مجموعه اسناد منبع به زبان مقصد خواهند بود و برای مشابهت‌یابی نیاز است از روش‌های بین‌زبانی کمک گرفته شود [18]، [19]، [20].

سیستم پیشنهادی در این مقاله مبتنی بر توسعه سیستم‌های نوع اول می‌باشد. اکثر سیستم‌های نوع اول بر مبنای یک چارچوب معین توسعه داده می‌شوند. چارچوب کلی سیستم‌های نوع اول شامل سه مرحله است:

- **بازیابی اسناد کاندیدا**

سیستم‌های کشف دستبرد علمی، وقتی در کاربردهای واقعی و با حجم داده بزرگ مورد استفاده قرار می‌گیرند، اعمال یک مرحله پیش‌پردازش با هدف محدود کردن اسناد منبع، برای آنها لازم و ضروری خواهد بود؛ بنابراین اولین گام، بازیابی اسناد مرتبط با سند پرس‌وجو است.

- **تشخیص مرز کپی‌برداری**

گام بعدی در سیستم‌های نوع اول، تشخیص مرز بخش‌هایی از سند پرس‌وجو و بخش‌هایی از اسناد منبع خواهد بود که با هم شباهت دارند. در این مرحله مشخص خواهد شد کدام بخش‌ها از هر یک از اسناد بازیابی شده، به کدام بخش‌ها از سند پرس‌وجو شباهت دارند و میزان این شباهت چقدر است.

- **پردازش نهایی به منظور آشکارسازی دستبرد علمی**

در این مرحله اگر در سند پرس‌وجو کپی‌برداری‌هایی تشخیص داده شود، اسناد منبعی که بوسیله سند پرس‌وجو کپی‌برداری شده‌اند، به ترتیب میزان شباهت‌شان به سند پرس‌وجو، رتبه‌بندی می‌شوند.

Stamatatos [3] کلیه روش‌های مشابهت‌یابی اسناد، را به دو دسته تقسیم می‌کند که عبارتند از روش‌های مبتنی بر بازیابی اطلاعات و روش‌های مبتنی بر اثرانگشت[۲]. در زمینه مشابهت‌یابی‌های مبتنی بر اثرانگشت، کارهای زیادی توسط محققان انجام گرفته است. کامران و همکارانش[1] نتایج حاصل از اعمال برخی از معروف‌ترین روش‌های اثرانگشت را در مجموعه‌ای از اسناد فارسی بررسی کردند.

یکی از روش‌های کشف دستبرد علمی مبتنی بر بازیابی اطلاعات که مورد توجه برخی محققین قرار گرفت، روش Rahman و همکارانش [5] است. آنها در روش خود برای مشابهت‌یابی اسناد از یک شبکه عصبی چند لایه به نام MLSOM [۳] استفاده کردند. در این سیستم ابتدا هر سند با یک ساختار داده درختی (سند–صفحه–پاراگراف) بازنمایی می‌شود. بر اساس ساختار داده درختی اسناد، از یک شبکه عصبی SOM سه لایه نیز استفاده شده است که لایه اول با پاراگراف‌های اسناد پایگاه داده آموزش می‌بیند و لایه دوم با صفحات اسناد و لایه سوم با کل اسناد پایگاه داده آموزش می‌بینند. سپس سند پرس‌وجو به کمک این شبکه عصبی در سه سطح پاراگراف، صفحه و سند مورد مشابهت‌یابی با اسناد منبع قرار می‌گیرد. این روش به دلیل هزینه محاسباتی کمی که در زمان اجرا دارد، قابل استفاده در پایگاه‌داده‌های بزرگ خواهد بود.

از جمله کسانی که سعی کردند MLSOM را بهبود دهند Zhang و همکارانش [6] بودند. آنها بر اساس این نکته که بهترین سطح مقایسه برای کشف دستبرد علمی، سطح جمله است، ساختار داده درختی پیشنهادی در [5] را به (سند– پاراگراف– جمله) تغییر دادند. همچنین برای کاهش زمان پیش‌پردازش معیاری را برای مشابهت‌یابی اسناد، جایگزین شبکه عصبی کردند اما ایراد این روش را می‌توان در شیوه بازنمایی جملات دانست. آنها یک مجموعه واژگان بزرگ را در نظر گرفته و هر جمله را با بردار حضور یا عدم حضور کلمات در این مجموعه واژگان بازنمایی می‌کند. در این روش، باید مجموعه واژگان بسیار بزرگ در نظر گرفته شود که این امر زمان محاسباتی را تحت تأثیر قرار می‌دهد و اگر مجموعه واژگان به اندازه کافی بزرگ نباشد، در عمل بسیاری از جملات، قابل بازنمایی نیستند زیرا کلمه‌ای از مجموعه واژگان در آنها حضور نخواهد داشت. یکی دیگر از روش‌های مبتنی بازیابی اطلاعات، روش Tiakas و همکارانش [7] است که آن را MSIDX نامیدند. روش MSIDX یک روش بازیابی اطلاعات مبتنی بر بردار است. ایده اصلی این است که بعد از ایجاد بردارهای توصیفگر از یک پایگاه داده، می‌توان کاردینالیتی هر بُعد در این بردارها را محاسبه کرد. کاردینالیتی یک بُعد، تعداد مقادیر مختلفی است که بُعد مربوطه در این پایگاه داده داشته است. بنابراین ابعادی که کاردینالیتی کمی دارند از اهمیت کمتری برخوردارند و ابعاد با کاردینالیتی بالا بیانگر تمایز بین اسناد هستند. بر این اساس ابعاد به نسبت مقدار کاردینالیتی‌شان اولویت‌بندی می‌شوند سپس بردارهای اسناد پایگاه‌داده بر اساس الگوریتم‌های مرتب‌سازی، مرتب می‌شوند. این روش از سرعت بسیار بالایی برخوردار است و قادر است روی پایگاه داده‌های بسیار بزرگ اجرا شود اما وقتی تعداد اسناد بازیابی شده زیاد باشد، دقت این روش بسیار کاهش می‌یابد و در واقع این روش فقط برای بازیابی اسناد با تعداد محدود، از دقت کافی برخوردار خواهد بود.

به طور کلی روش‌های مبتنی بر بازیابی اطلاعات در کشف کپی‌برداری‌های هوشمندانه، موفق‌تر هستند و سیستم‌های مبتنی بر اثرانگشت اگر چه در کشف کپی‌برداری‌های دقیق کارا هستند اما قابل توسعه برای کشف انواع تقلب‌های هوشمندانه نیستند و بر همین اساس در روش پیشنهادی تلاش بر این بوده است که از روش‌های مبتنی بازیابی اطلاعات استفاده شود به طوری که بتوان آن را برای دستبردهای هوشمندانه نیز توسعه داد. سیستم پیشنهادی قادر به کشف انواع کپی‌برداری‌های دقیق، کپی‌برداری با تغییرات و برخی کپی‌برداری‌های هوشمندانه مانند خلاصه‌سازی (کاهش، تقسیم و تلفیق جملات) و جایگزینی کلمات با مترادف‌هایشان می‌باشد. همچنین، سیستم پیشنهادی در رویکردی جدید تصاویر اسناد را نیز مورد بررسی قرار می‌دهد به طوری که در رتبه‌بندی اسناد منبع برای هر سند پرس‌وجو، علاوه بر متن اسناد، تصاویر آنها نیز نقش خواهند داشت.

2- الگوریتم پیشنهادی

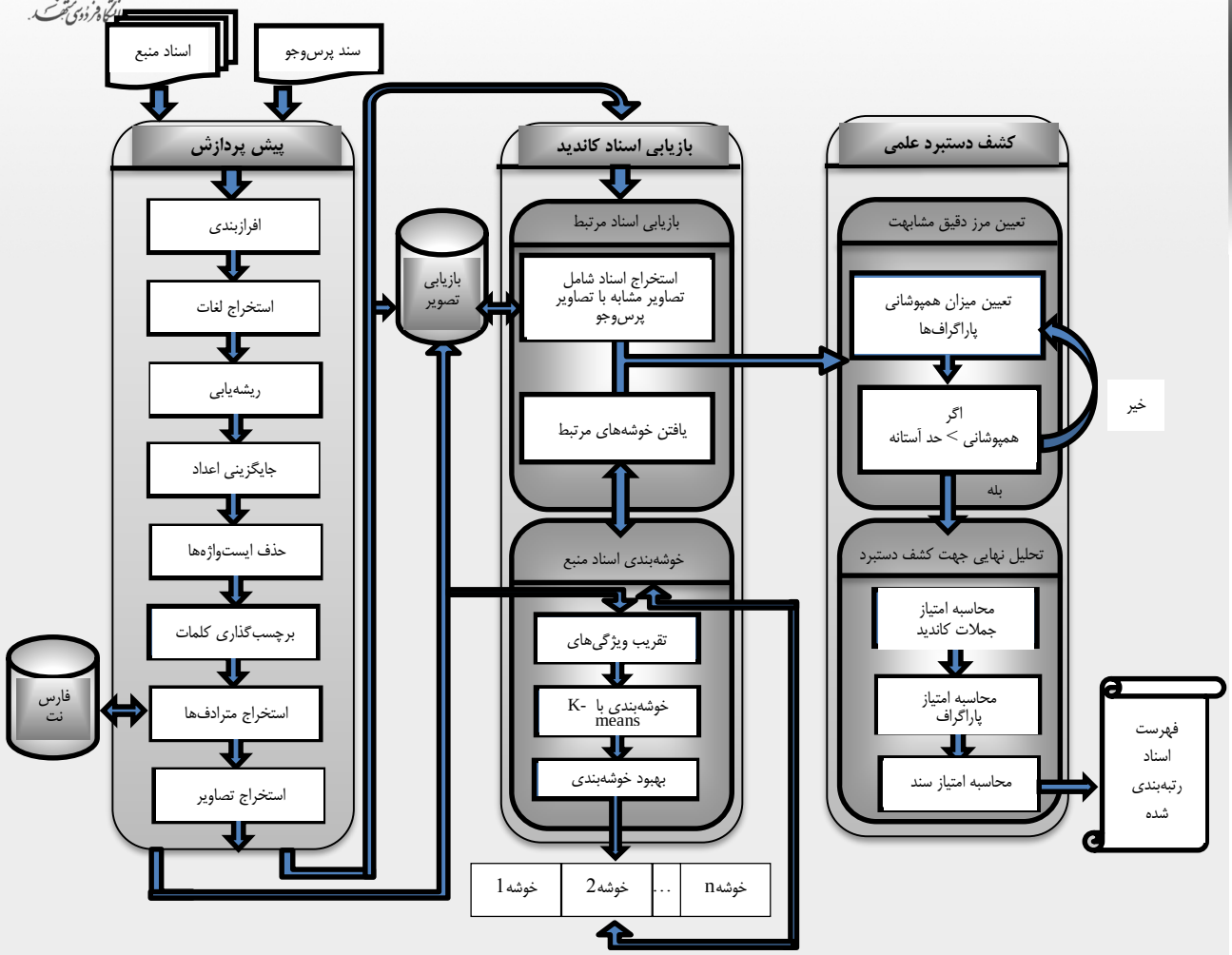
ما روش پیشنهادی برای کشف دستبرد علمی را MLO [۴] نامیده‌ایم که نحوه عملکرد آن در شکل (1) نشان داده شده است. این سیستم از سه بخش کلی پیش‌پردازش، بازیابی اسناد کاندیدا و کشف دستبرد علمی تشکیل می‌شود که در ادامه هر یک از این مراحل شرح داده خواهند شد.

2-1 پیش‌پردازش

طبق شکل (1) پیش‌پردازش‌هایی که باید روی اسناد صورت گیرند به ترتیب زیر خواهند بود:

- **افراز[۵] اسناد**

در MLO ما اسناد را با ساختار داده درختی در سه سطح (سند– پاراگراف– جمله)، بازنمایی می‌کنیم. برای به کار گیری ساختار داده درختی لازم است اسناد به بخش‌های کوچکتر افراز شوند. تلاش بر این است که در افراز بندی، بخش‌هایی که بوسیله نویسنده متن، مرتبط تشخیص داده شده‌اند در یک افراز قرار بگیرند. بر این اساس بخش‌بندی اسناد به پاراگراف‌هایش، یک بخش‌بندی متداول و مفید است. نکته دیگری که باید به آن توجه کرد این است که هدف از افرازبندی اسناد، تقسیم سند به واحدهای کوچکتر است به



شکل (1): چارچوب سیستم کشف دستبرد علمی پیشنهادی

طوری که در مراحل بعد، توزیع مکانی ویژگی‌ها را در هر یک از این واحدها مورد بررسی قرار می‌دهیم بنابراین از ایجاد واحدهای خیلی کوتاه جلوگیری می‌کنیم. برای این منظور ابتدا یک سند به پاراگراف‌هایش شکسته می‌شود سپس تا زمانی که تعداد کلمات یک پاراگراف کمتر از یک حد آستانه باشد، با پاراگراف‌های بعدی ادغام خواهد شد. پس از بخش‌بندی سند به پاراگراف‌ها، جملات از هر پاراگراف استخراج می‌شود.

- **استخراج لغات**

ما برای استخراج کلمات یک متن، از ابزار STEP-1 [8] استفاده کرده‌ایم.

- **ریشه‌یابی**

برای ریشه‌یابی کلمات نیز از STEP-1 کمک گرفتیم و ریشه تصریفی هرکلمه را استخراج کردیم.

- **جایگزینی اعداد**

در این مرحله تمام اعداد موجود در متن با یک نشانه خاص بازنمایی می‌شوند.

- **حذف ایست‌واژه‌ها**

Davarpanah و همکارانش [9] فهرستی از ایست‌واژه‌های زبان فارسی را ارائه داده‌اند و ما نیز ایست‌واژه‌ها را بر اساس این فهرست شناسایی کرده و حذف می‌کنیم.

- **تعیین مقوله نحوی کلمات**

مقوله نحوی تمام کلمات هر پاراگراف، مشخص می‌شود و از بین آنها فقط اسامی، صفات و کلمات انگلیسی استخراج می‌شوند.

- **استخراج مترادف‌های کلمات در هر جمله**

این مرحله از پیش‌پردازش، فقط برای اسناد منبع انجام می‌شود، به این ترتیب که برای تمام کلمات در هر جمله، مترادف‌های آنها به کمک ابزار فارس‌نت[10] استخراج می‌شود.

- **استخراج تصاویر از اسناد**

تمام تصاویر موجود در سند، در این مرحله استخراج می‌شوند.

2-2 بازیابی اسناد کاندیدا

همان‌طور که بیان شد، در سیستم‌های نوع اول برای هر سند پرس‌وجو، اسناد مرتبط بازیابی می‌شوند. بازیابی اسناد، می‌تواند به دو شیوه صورت گیرد. در شیوه اول، میزان شباهت هر یک از اسناد منبع به سند پرس‌وجو محاسبه شده و اسنادی که بیشترین شباهت را به سند پرس‌وجو دارند، بازیابی می‌شوند [23]، [24]. این مشابهت‌سنجی می‌تواند حجم محاسباتی زیادی را به سیستم تحمیل کند در حالی که هدف از بازیابی اسناد کاندیدا، ممانعت از مقایسه دقیق سند پرس‌وجو با کل اسناد منبع است. شیوه دیگر، استفاده از تکنیک‌های خوشه‌بندی می‌باشد. در این روش اسناد منبع بر اساس میزان شباهتشان خوشه‌بندی می‌شوند و شباهت‌سنجی سند پرس‌وجو تنها با ویژگی‌های خوشه‌ها صورت می‌گیرد و اسناد خوشه‌ای که سند پرس‌وجو، بیشترین شباهت را با ویژگی‌های آن داشته باشد، به عنوان اسناد کاندیدا در نظر گرفته می‌شوند[25]،[26]. ما نیز تکنیک‌های خوشه‌بندی را به دلیل

سرعت خوبی که می‌توانند در بازیابی اسناد داشته باشند، در MLO به کار برده‌ایم بنابراین مرحله بازیابی اسناد کاندیدا در MLO شامل دو بخش خواهد بود بخش اول، خوشه‌بندی اسناد پایگاه‌داده است و بخش دوم، مرحله بازیابی اسناد مرتبط با سند پرس‌وجو است.

2-2-1 خوشه‌بندی اسناد پایگاه‌داده

روش پیشنهادی برای خوشه‌بندی اسناد پایگاه‌داده به این ترتیب عمل می‌کند که ابتدا ویژگی‌های خوشه‌ها حدس زده می‌شوند و سپس اسناد بر اساس دارا بودن این ویژگی‌ها، خوشه‌بندی شده و در آخر پردازشی با هدف بهبود خوشه‌بندی صورت می‌گیرد. جزئیات هر یک از مراحل به این شرح است:

- **استخراج ویژگی‌های خوشه‌ها**

برای استخراج ویژگی‌ها، ابتدا تمام لغات اسناد منبع استخراج خواهند شد و هر لغت بر اساس تعداد تکرارش در مجموعه اسناد منبع و تعداد اسنادی که شامل این لغت هستند، در جدولی که توسط شکل(2) نشان داده شده است، قرار می‌گیرد. بر اساس نظر Makrehchi و همکارانش[11] تمامی کلمات استخراج شده از یک مجموعه سند، می‌توانند به سه دسته ایست‌واژه‌ها، کلمات کلیدی و ویژگی‌ها تقسیم شوند. ایده کلی این تقسیم‌بندی بر این اساس است که کلماتی که در کلیه اسناد کلیه کلاس‌ها پخش شده باشند، دارای ارزش و مفهوم کمتری هستند و به عنوان ایست‌واژه‌ها به شمار می‌روند و کلماتی که فقط در تعدادی از اسناد فراوانی زیادی داشته باشند درحالی که در سایر اسناد فراوانی کمتری دارند، احتمالاً از ویژگی‌های آن اسناد خواهند بود و لغاتی که در تعداد بسیار کمی از اسناد فراوانی زیادی دارند و در بقیه اسناد فراوانی کمی دارند، عموماً کلمات کلیدی آن اسناد به شمار می‌روند. بنابراین تمام کلمات بر اساس فراوانی آنها و تعداد اسنادی که آنها را شامل می‌شوند در این جدول دسته‌بندی خواهند شد. هر کدام از مقادیر حد آستانه در این جدول کاملا بستگی به بزرگی مجموعه اسناد منبع دارد. به این ترتیب ویژگی‌ها تقریب زده می‌شوند. بعد از استخراج ویژگی‌ها از بین کلمات، بردار هیستوگرام تمام اسناد منبع بر اساس ویژگی‌های استخراج شده محاسبه می‌شود، به این ترتیب که برای هر سند یک بردار n بیتی (n تعداد ویژگی‌های استخراج شده می‌باشد) به نام $V(v_1, v_2, \dots, v_n)$ در نظر گرفته می‌شود. به طوری که اگر ویژگی iام در سند حضور داشته باشد مقدار v_i یک و اگر سند شامل این ویژگی نباشد v_i صفر خواهد بود.

- **خوشه‌بندی اسناد**

ما از الگوریتم K-means برای خوشه‌بندی اسناد پایگاه‌داده استفاده کرده‌ایم، به این ترتیب که بردارهای $V(v_1, v_2, \dots, v_n)$ حاصل از مرحله قبل، داده‌های ورودی الگوریتم K-means هستند که آنها را خوشه‌بندی می‌کنیم. در الگوریتم K-means از معیار شباهت کسینوسی، برای مشابهت‌سنجی بردارها استفاده می‌شود و سند به خوشه‌ای که بیشترین شباهت را با مرکز آن خوشه دارد، تعلق می‌گیرد. این کار ادامه می‌یابد تا زمانی که مراکز خوشه‌ها تغییر نکنند.

- **بهبود خوشه‌بندی**

یک روش بهبود خوشه‌بندی، بهبود ویژگی‌های استخراج شده است. ما برای استخراج ویژگی‌های هر یک از خوشه‌ها به این ترتیب عمل خواهیم کرد که تمام بردارهای $V(v_1, v_2, \dots, v_n)$ برای اسناد موجود در این خوشه با هم جمع خواهند شد و مقادیر حاصل، در یک بردار n بیتی به نام $H(h_1, h_2, \dots, h_n)$ ذخیره می‌شوند. به طوری که h_1 تعداد اسنادی را در این خوشه نشان می‌دهد که اولین ویژگی از ویژگی‌های استخراج شده در مرحله اول را دارا هستند. بردار H برای تمام خوشه‌ها محاسبه می‌شود. حال بر اساس این بردارها، می‌توان وزن هر ویژگی را در خوشه مربوطه طبق رابطه (1) محاسبه کرد.

$$w_{h_1} = \frac{f_d}{f_c} \qquad (1)$$

که در آن f_d بیانگر تعداد اسنادی از خوشه مربوطه هستند که شامل ویژگی اول بوده و f_c بیانگر تعداد خوشه‌هایی است که این ویژگی را دارند. به عبارت دیگر، ویژگی‌های یک خوشه بر اساس اینکه چند سند در آن خوشه، این ویژگی را دارا هستند نسبت به تعداد خوشه‌هایی که این ویژگی را شامل می‌شوند، وزن‌دهی می‌شوند؛ اکنون ویژگی‌ها بر اساس وزن‌شان به صورت نزولی مرتب شده و n لغت اول از هر خوشه به عنوان ویژگی‌های آن خوشه در نظر گرفته می‌شوند. به همین ترتیب ویژگی‌ها برای تمام خوشه‌ها به دست خواهند آمد. سپس برای هر یک از اسناد منبع، درجه تعلق سند به هر کدام از خوشه‌ها محاسبه شده و اسناد در خوشه‌ای قرار می‌گیرند که بیشترین درجه تعلق را به آن دارند. عدد n بستگی به بزرگی مجموعه اسناد منبع دارد که ما به طور تجربی برای تعداد ویژگی‌های استخراج شده از خوشه‌ها که به طور میانگین حدود 300 ویژگی بوده است، عدد n را 50 در نظر گرفته‌ایم. درجه تعلق به خوشه‌ها برای هر سند بر اساس رابطه (2) محاسبه می‌شود:

$$Membership = \frac{Sum \times Count}{\sqrt{T_D}} \qquad (2)$$

که در آن Sum جمع تعداد تکرار تمام ویژگی‌های این خوشه در سند مربوطه است، $Count$ تعداد ویژگی‌هایی از این خوشه که در سند مربوطه حضور داشته‌اند و T_D که طول سند (تعداد لغات موجود در سند) می‌باشد. از آنجا که پایگاه‌داده شامل اسنادی با طول‌های متفاوت است، درجه تعلق اسناد نسبت به طول آنها نرمالایز می‌شوند.

2-2-2 بازیابی اسناد مرتبط با سند پرس‌وجو

در سیستم پیشنهادی، بازیابی اسناد کاندیدا به صورت زیر خواهد بود:

الف- ابتدا درجه تعلق سند پرس‌وجو به هر یک از خوشه‌ها بر اساس رابطه (2) محاسبه می‌شود و اسناد خوشه‌ای که سند پرس‌وجو بیشترین درجه تعلق را به آن دارد، به عنوان اسناد کاندیدا بازیابی می‌شوند.

ب- چنانچه سند پرس‌وجو به خوشه‌های دیگر درجه تعلق بیش از یک حد آستانه داشته باشد، اسناد آن خوشه‌ها نیز بازیابی می‌شوند. حد آستانه برای هر خوشه، میانگین درجه تعلق‌هایی است که اسناد آن خوشه به خوشه مربوطه دارند.

ج- در MLO، اسنادی از منبع که شامل هر یک از تصاویر سند پرس‌وجو باشند نیز بازیابی می‌شوند. این کار بر اساس این ایده انجام می‌شود که اگر از تصویر یک سند، کپی‌برداری شده باشد احتمال آن وجود دارد که از متن آن نیز کپی‌برداری صورت گرفته باشد؛ بنابراین در MLO از یک سیستم بازیابی تصویر نیز کمک گرفته می‌شود و اگر سندی در منبع وجود داشته باشد که تصویری (تصاویری) مشابه با سند پرس‌وجو دارد، توسط سیستم بازیابی می‌شود. به این ترتیب مجموعه اسناد کاندیدا که از مراحل (الف)، (ب) و (ج) حاصل می‌شوند، به عنوان اسناد مرتبط با سند پرس‌وجو در نظر گرفته می‌شوند؛ سپس برای رتبه‌بندی این اسناد از بردارهای حضور یا عدم حضور ویژگی‌های تقریب زده شده استفاده می‌کنیم و شباهت کسینوسی برای بردار

سند پرس‌وجو با هر یک از بردارهای اسناد بازیابی شده بدست خواهد آمد. از اسناد بازیابی شده جمع می‌شود. بنابراین اسناد خوشه‌ای که درجه تعلق سند پرس‌وجو به آن بیشتر است، امتیاز بیشتری خواهند داشت و در صورتی که همه اسناد بازیابی شده متعلق به یک خوشه باشند یک مقدار ثابت به امتیاز همه اسناد افزوده می‌شود که در رتبه‌بندی بی‌تأثیر خواهد بود. از بین این اسناد N سند با رتبه برتر به مرحله بعد وارد می‌شوند.

تعداد اسنادی که شامل کلمه هستند

تعداد تکرار کلمه در اسناد منبع	زیاد	متوسط	کم
زیاد	ایست‌واژه	ویژگی	کلمه کلیدی
متوسط	ایست‌واژه	ویژگی	کلمه‌کلیدی
کم	ایست‌واژه	ایست‌واژه	ایست‌واژه

شکل(2): دسته‌بندی کلمات پایگاه‌داده

3-2 کشف دستبرد علمی

این مرحله از دو بخش تشکیل می‌شود. بخش اول مزر مشابهت‌ها را بین دو سند مشخص کرده و بخش دوم رتبه‌بندی اسناد منبع است.

- **تعیین مرز دقیق مشابهت‌ها**

پس از بازیابی اسناد کاندیدا، لازم است در هر سند کاندیدا، قسمت‌هایی که مشابهتی به سند پرس‌وجو دارند، مشخص شوند. برای این منظور ما در این مرحله یک مشابهت‌یابی در سطح پاراگراف‌ها انجام می‌دهیم و پاراگراف‌هایی که میزان شباهت‌شان با پاراگراف مورد نظر از یک حد آستانه بیشتر باشد، بازیابی شده و به مرحله بعد وارد می‌شوند. فرض بر این است که در متون تخصصی، از بین کلمات یک پاراگراف اسامی، صفات و کلماتِ انگلیسی (کلمات اختصاری که در متن فارسی نیز به کار می‌روند) بیشترین نقش را در تعیین موضوع آن پاراگراف دارند و دیگر مقوله‌های نحوی برای این منظور عموماً از اهمیت کمتری برخوردارند. بنابراین ما برای کاهش محاسبات، یک پاراگراف را تنها با اسامی، صفات و کلمات انگلیسی موجود در پاراگراف، بازنمایی می‌کنیم. سپس برای تعیین شباهت دو پاراگراف، میزان همپوشانی دو پاراگراف بر اساس معیار شباهت Jaccard طبق رابطه (3) محاسبه می‌شود:

$$\delta(x,y) = \frac{|x \cap y|}{|x \cup y|} \qquad (3)$$

که در آن x و y دو مجموعه هستند و $\delta(x,y)$ میزان همپوشانی دو مجموعه را بر اساس نسبت اشتراک به اجتماع دو مجموعه، تعیین می‌کند. برای هر پاراگراف از سند پرس‌وجو میزان شباهت آن با تمام پاراگراف‌های سند کاندیدا محاسبه شده و اسناد کاندیدا که حداقل یک پاراگراف مشابه با سند پرس‌وجو دارند به مرحله بعد وارد می‌شوند.

- **تجزیه و تحلیل نهایی برای کشف دستبرد علمی**

آخرین مرحله، تحلیل نهایی بخش‌های مشابه بین سند پرس‌وجو و اسناد کاندیدا است. خروجی این مرحله فهرستی از اسناد منبع است که بر اساس مقدار شباهت کسینوسی با مقدار درجه تعلق سند پرس‌وجو به خوشه هر یک میزان شباهتشان به سند پرس‌وجو رتبه‌بندی شده‌اند. رتبه‌بندی اسناد کاندیدا بر اساس مشابهت‌یابی در سطح جملات به صورت زیر خواهد بود که برای محاسبه شباهت بین یک جمله از سند پرس‌وجو و یک جمله از سند کاندیدا، میزان همپوشانی مجموعه کلمات جمله پرس‌وجو را یک بار با مجموعه کلمات یک جمله کاندیدا و یک بار هم با مجموعه کلمات مترادف آن جمله محاسبه می‌شود و جمع این دو همپوشانی، میزان شباهت دو جمله را مشخص می‌کند که بر اساس رابطه (4) محاسبه می‌شود:

$$Sim(S_{iQ}, S_{js}) = \delta(S_{iQ}, S_{js}) + W \times \delta(S_{iQ}, SYN_{js}) \qquad (4)$$

در رابطه (4) S_{iQ} جمله iام از سند پرس‌وجو و S_{js} جمله jام از سند کاندیدا است. $\delta(S_{iQ}, S_{js})$ معیار شباهت Jaccard است که بر اساس رابطه (3) محاسبه می‌شود. SYN_{js} مجموعه کلمات مترادف استخراج شده برای جمله jام از سند کاندیدا است و W وزن را نشان می‌دهد که اگر یک در نظر گرفته شود کپی‌برداری دقیق از یک متن با کپی‌برداری از آن با جایگزینی کلمات، یکسان در نظر گرفته می‌شود و اگر W عددی کمتر از یک باشد به کپی‌برداری دقیق نسبت به کپی‌برداری با جایگزینی کلمات، وزن بیشتری نسبت داده می‌شود. نتایج بررسی تأثیر پارامتر W در جدول (3) آمده است. از بین تمام امتیازهای دریافتی توسط یک جمله از سند کاندیدا، ماکزیمم امتیاز، به عنوان امتیاز آن جمله در نظر گرفته می‌شود. بعد از مشخص شدن امتیاز تمام جملات پاراگراف کاندیدا، امتیاز پاراگراف‌ها یعنی $PSC(P_i)$، از طریق رابطه (5) محاسبه می‌شود. همان‌طور که در رابطه مشخص شده است، امتیاز یک پاراگراف از جمع امتیاز جملات آن پاراگراف به نسبت تعداد جملات پاراگراف ذکر شده حاصل می‌شود.

$$PSC(p_i) = \frac{\sum_{i=1}^{m} SSC(s_i)}{m} \qquad (5)$$

که در آن $SSC(s_i)$ امتیاز جمله iام در پاراگراف p_i و m تعداد جملات پاراگراف p_i می‌باشد و در نهایت، امتیاز هر سند کاندیدا طبق رابطه (6) از جمع امتیاز پاراگراف‌هایش به نسبت تعداد پاراگراف‌های آن سند حاصل می‌شود و همچنین علاوه بر امتیازی که یک سند از مشابهت‌یابی متن خود دریافت می‌کند، امتیاز تصاویر سند کاندیدا نیز به امتیاز سند افزوده می‌شود. امتیازی که یک سند کاندیدا از هر تصویرش دریافت می‌کند از رابطه (7) بدست می‌آید که در آن یک سند به ازای هر تصویر مشابه با تصاویر سند پرس‌وجو یک امتیاز ثابت k را دریافت می‌کند.

$$DSC = \frac{\sum_{i=1}^{n} PSC(P_i)}{n} + \sum_{j=1}^{l} ISC(img_j) \qquad (6)$$

در رابطه (6) امتیاز سند کاندیدا و DSC امتیاز پاراگراف iام از سند کاندیدا و n تعداد پاراگراف‌های سند کاندیدا می‌باشد. $ISC(img_j)$ امتیاز تصویر jام از تصاویر سند کاندیدا است که از رابطه (7) به دست می‌آید و l تعداد تصاویر سند کاندیدا می‌باشد.

$$ISC(img_j) = \begin{cases} k, & \text{اگر } Img_j \text{ مشابه هر یک از تصاویر سند پرس‌وجو هست} \\ 0, & \text{در غیر اینصورت} \end{cases} \qquad (7)$$

با استفاده از این شیوه افراز اسناد، اگر اسناد با حجم بالا در مجموعه اسناد منبع حضور داشته باشند، چنانچه حداقل شامل یک پاراگراف باشند که امتیاز بیش از حدآستانه کسب کند، سند مربوطه فارغ از حجم کل سند، در فهرست اسناد رتبه‌بندی شده قرار می‌گیرند. همچنین در پیاده‌سازی سیستم پیشنهادی

تمام پاراگراف‌هایی که حداقل امتیاز را کسب کرده باشند به عنوان پاراگراف‌های مشکوک به کاربر نمایش داده می‌شوند. در مواردی که جمع امتیاز جملات مشکوک از یک پاراگراف از حد آستانه کمتر باشد، سیستم اعلام کشف دستبرد علمی نخواهد کرد که این موارد را با تنظیم حد آستانه می‌توان مدیریت کرد که با مشابهت چند درصد از جملات یک پاراگراف، می‌توان گفت پاراگراف مربوطه دچار دستبرد علمی شده است؟ اگر چه حد آستانه را می‌توان تا حدی کاهش داد که به ازای هر جمله مشابه پاراگراف مربوطه مشکوک در نظر گرفته شود اما باید توجه کرد که اعمال محدودیت‌های مناسب برای ورود اسناد به مرحله مشابهت‌یابی در سطح جمله باعث می‌شود سیستم قادر به انجام پردازش‌های وسیع‌تری در سطح جمله باشد که منجر به کشف انواع مختلف تقلب شود. بنابراین با برقراری این توازن می‌توان به نتایج مطلوب‌تری رسید.

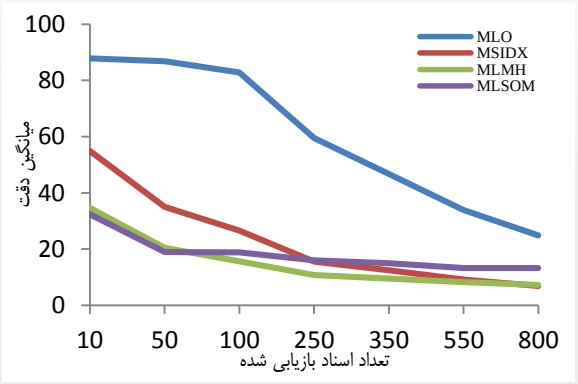
4-نتایج

مجموعه پایگاه داده‌ای که برای این سیستم تهیه شده است شامل پنج هزار سند بوده که حدود یک هزار سند آن، از مجموعه مقالات انجمن کامپیوتر ایران در دوره‌های نهم تا شانزدهم، 200 مقاله از برخی کنفرانس‌های داخلی در حوزه کامپیوتر، هزار سند از مجموعه مقالات و پایان‌نامه‌های آماده در حوزه کامپیوتر که در فروشگاه‌های اینترنتی خرید و فروش می‌شود و بقیه اسناد از مجموعه مقالات ویکی‌پدیا در حوزه کامپیوتر هستند. کمترین طول اسناد در این مجموعه، یک پاراگراف و بیشترین طول آنها 326 صفحه است. اسناد این مجموعه، به زبان فارسی می‌باشند. برای تهیه اسناد پرس‌وجو، از میان پنج هزار سند منبع، 100 سند منبع را به طور تصادفی جهت کپی برداری انتخاب کرده و تغییراتی مانند حذف و افزودن جملات، جابجایی جملات، حذف و افزودن برخی کلمات از جملات، جایگزینی برخی کلمات با مترادف‌هایشان و ترکیب برخی از بخش‌های کپی‌شده از اسناد مختلف را در آنها اعمال کرده‌ایم. آزمایش‌های انجام شده در این بخش شامل بررسی عملکرد سیستم بازیابی اطلاعات پیشنهادی و بررسی عملکرد سیستم کشف دستبرد علمی پیشنهادی می‌باشد. در یک آزمایش، برای بررسی عملکرد سیستم بازیابی اطلاعات، میانگین دقت و میانگین فراخوانی بازیابی اسناد، برای روش MLO و سه روش MSIDX، MLSOM و MLMH محاسبه شده است. نتایج این آزمایش در جدول (1) و همچنین نمودار شکل (3) آمده است. دقت بازیابی اسناد از نسبت تعداد اسناد صحیح بازیابی شده به تعداد اسناد بازیابی شده بدست می‌آید و فراخوان از نسبت تعداد اسناد صحیح بازیابی شده به تعداد کل اسناد مرتبط در پایگاه‌داده بدست می‌آید. همانطور که از نتایج مشاهده می‌شود الگوریتم پیشنهادی MLO با بازیابی اسناد مبتنی بر خوشه‌بندی از روش‌های بازیابی مستقیم اسناد، عملکرد بهتری دارد. بازیابی مبتنی بر خوشه‌بندی علاوه بر اینکه از نظر دقت و فراخوانی کارایی خوشه بندی خوشه‌بندی علاوه بر اینکه از نظر دقت و فراخوان کارایی مناسبی دارد، به دلیل اینکه خوشه‌بندی اسناد منبع به عنوان یک پیش پردازش صورت می‌گیرد، زمان پاسخگویی سیستم نسبت به روش‌های بازیابی مستقیم کوتاهتر خواهد بود و همین امر سبب می‌شود سیستم کشف دستبرد علمی قابل استفاده برای پایگاه‌داده‌های واقعی باشد. باید توجه داشت که بازیابی اسناد نباید منجر به از دست دادن اسناد مرتبط شود زیرا در این صورت احتمال شکست در کشف دستبرد علمی افزایش می‌یابد. برای بررسی عملکرد روش‌های بازیابی مورد بحث، در از دست دادن سند منبع مورد نظر، آزمایش حضور سند منبع در فهرست اسناد بازیابی شده صورت گرفت. نتایج حاصل از این آزمایش در شکل (4) آمده است. نتایج نشان می‌دهد در روش MLO میانگین حضور سند منبع نسبت به روش‌های دیگر بسیار بالاتر است. نکته مهم دیگر ثابت ماندن میانگین حضور سند منبع برای مقادیر مختلف از تعداد اسناد بازیابی شده، در روش MLO می‌باشد. همانطور که مشاهده می‌شود وقتی تعداد اسناد بازیابی شده بیشتر از 5 سند باشد نتایج برای حضور سند منبع در بیشترین حد ثابت خواهد ماند. این نشان می‌دهد معیار رتبه‌بندی اسناد در خوشه بازیابی شده برای سند پرس‌وجو، یک معیار مناسب است و چنانچه در بازیابی N سند، عدد N بزرگتر از 5 در نظر گرفته شود، این محدود کردن اسناد بازیابی شده هیچ تأثیر منفی در کارایی سیستم کشف دستبرد علمی نخواهد داشت.

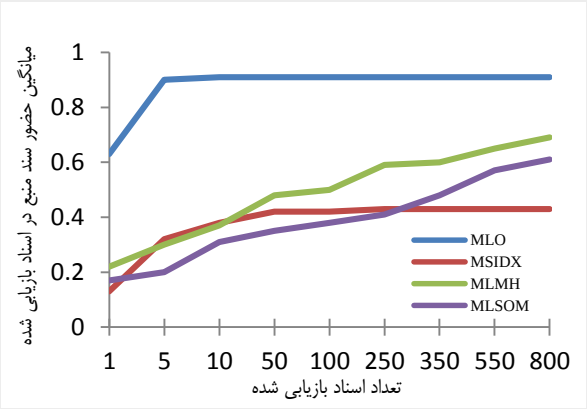
در حال حاضر الگوریتم‌های مبتنی بر اثرانگشت در کشف کپی‌برداری‌های تحت‌اللفظی موفق‌ترین روش‌ها محسوب می‌شوند[3]. بر همین اساس برای ارزیابی عملکرد سیستم کشف دستبرد علمی پیشنهادی، ما روش پیشنهادی را با یکی از روش‌های مبتنی بر اثر انگشت به نام روش SimHash مقایسه کرده‌ایم. دلیل انتخاب روش SimHash از بین روش‌های مبتنی بر اثرانگشت نیز گزارش کامران و همکارانش[1] می‌باشد که روش SimHash را مناسب برای اعمال بر مجموعه‌ای از اسناد فارسی اعلام کردند. نتایج حاصل از این مقایسه در جدول (2) آمده است. در این جدول میانگین رتبه، میانگین رتبه‌های سند منبع برای صد سند پرس‌وجو است و مینیمم رتبه، بهترین رتبه‌ای است که سند کپی‌برداری شده، دریافت کرده است و ماکزیمم رتبه، بدترین رتبه‌ای است که سند کپی‌برداری شده در این آزمایش‌ها دریافت کرده است. همچنین منظور از تعداد شکست‌ها، تعداد دفعاتی است که سند منبع مورد نظر در فهرست اسناد رتبه‌بندی شده یافت نمی‌شود. از نتایج بدست آمده مشخص است که میانگین رتبه برای سند منبع کپی‌برداری شده در روش MLO بسیار کمتر از روش SimHash است و این نشان می‌دهد که MLO در کشف سند کپی‌برداری شده عملکرد قابل قبولی دارد. نکته قابل توجه دیگر، تعداد شکست‌هاست. روش‌های مبتنی بر اثرانگشت در مشابهت‌یابی سرعت بالایی دارند بنابراین می‌توانند مقایسه سند پرس‌وجو را با همه اسناد منبع انجام دهند و نیاز به بازیابی اسناد کاندیدا ندارند و این منجر به کاهش تعداد شکست‌ها می‌شود در حالی که معمولا در روش‌های مبتنی بر بازیابی اطلاعات تعداد شکست‌ها زیاد است. اما علت استفاده از روش‌های مبتنی بر بازیابی اطلاعات، محدودیت‌های روش‌های مبتنی بر اثر انگشت در کشف کپی‌برداری‌های هوشمندانه است.

جدول (1): نتایج بازیابی اسناد کاندیدا برای صد سند پرس‌وجو

روشها	تعداد اسناد بازیابی شده							
	10	100	350	800	10	100	350	800
	میانگین دقت				میانگین فراخوانی			
MLO	87.8	82.8	46.7	24.9	6.4	54.5	82.4	88.4
MSIDX	54.94	26.58	12.5	6.94	3.61	14.2	19.4	22.5
MLMH	34.64	15.65	9.5	7.33	1.69	6.50	13.45	23.73
MLSOM	32.45	18.87	14.98	12.22	1.13	6.21	22.18	26.46



شکل (3): میانگین دقت برای صد سند پرس‌وجو



شکل(4): میانگین حضور سند منبع برای اسناد بازیابی شده

جدول (2): نتایج رتبه‌بندی برای صد سندپرس‌وجو

	نتایج رتبه‌بندی اسناد			
	میانگین رتبه‌ها	مینیمم رتبه	ماکزیمم رتبه	تعداد شکست‌ها
MLO	**5**	**1**	**78**	**8**
SimHash	**100**	**1**	**1534**	**0**

جدول(3): بررسی تأثیر پارامتر وزن w در مشابهت‌یابی بین جملات

مقدار پارامتر وزن w	میانگین حضور سند منبع مورد نظر در رتبه اول	میانگین اختلاف امتیاز سند منبع در کپی‌برداری دقیق و کپی‌برداری غیر مستقیم
0.3	٪99	27.83
0.4	٪100	25.31
0.5	٪100	22.94
0.7	٪100	19.69
0.8	٪100	17.26

پارامتر وزن w با دادن وزن بیشتر به مشابهت‌یابی بین کلمات سند پرس‌وجو و کلمات سند منبع و دادن وزن کمتر به مشابهت‌یابی بین کلمات سند پرس‌وجو و مترادف‌های سند منبع، سعی در نمایش این تفاوت دارد. نکته‌ای که باید به آن توجه کرد این است که این تفاوت ایجاد شده تا چه اندازه باشد که کارایی سیستم را کاهش ندهد یعنی هنگامی که سند پرس‌وجو از یک سند منبع غیر مستقیم و با جایگزینی کلمات کپی‌برداری می‌کند در عین حال که سند منبع امتیاز کمتری را نسبت به حالت دقیقاً کپی شده دریافت می‌کند، اما همچنان در فهرست رتبه‌بندی اسناد منبع در رتبه اول باقی بماند. برای یافتن بهینه، آزمایش‌هایی صورت گرفت که در این آزمایشات میانگین حضور سند منبع مورد نظر در رتبه اول برای هر مقدار از پارامتر w مورد بررسی قرار گرفته است. علاوه بر این در تمام موارد امتیازی که هر سند پرس‌وجو در هنگام کپی‌برداری غیر مستقیم کسب کرده است کمتر از زمانی است که همان سند پرس‌وجو از بخش موردنظر دقیقاً کپی‌برداری کرده است. میانگین اختلاف امتیازها برای کپی‌برداری دقیق و کپی‌برداری غیر مستقیم اسناد پرس‌وجو محاسبه شده است. بنابراین ما به دنبال مواردی هستیم که با وجود اختلاف امتیاز بین دو نوع کپی‌برداری، اسناد منبع مورد نظر با بالاترین رتبه بازیابی می‌شوند. ما مقدار 0.7 را برای وزن w انتخاب کردیم که مطمئن باشیم در کارایی سیستم تأثیر منفی نخواهیم داشت.

5- نتیجه‌گیری

الگوریتم پیشنهاد شده در این مقاله یک سیستم کشف دستبرد علمی مبتنی بر بازیابی اطلاعات است که قابل توسعه به کشف انواع کپی‌برداری‌های هوشمندانه است. در این الگوریتم از روش‌های بازیابی اطلاعات مبتنی بر خوشه‌بندی استفاده شده است که نتایج آزمایشات حاکی از عملکرد قابل قبول سیستم بازیابی اطلاعات پیشنهادی است به طوری که در ٪91 موارد، سند منبع مورد نظر در 5 سند بازیابی شده با رتبه برتر حضور دارد. بر اساس نتایج به دست آمده ما روش‌های بازیابی اطلاعات مبتنی بر خوشه‌بندی را مناسب‌تر از روش‌های بازیابی اطلاعات مستقیم، برای سیستم‌های کشف دستبرد علمی می‌دانیم. در این روش، خوشه‌بندی اسناد فقط یک بار در زمان پیش پردازش صورت می‌گیرد و هنگامی که سند پرس‌وجو وارد سیستم می‌شود تنها بر اساس ویژگی‌های هر دسته، میزان تعلق آن به دسته‌ها مشخص می‌شود و سند مربوطه در دسته‌ای که بیشترین میزان تعلق را دارد قرار می‌گیرد و لازم نیست که خوشه‌بندی جدیدی صورت گیرد. این شیوه خوشه‌بندی به دلیل اینکه بار محاسباتی زیادی را در زمان اجرا به سیستم وارد نمی‌کند سبب شده که سیستم پیشنهادی قابل توسعه برای پایگاه داده‌های با حجم بالا نیز باشد که این یکی از مزیت‌های روش پیشنهادی است. در روش کشف دستبرد علمی پیشنهادی در مرحله اول مقایسه با مجموعه اسناد منبع در سطح سند انجام می‌شود، از بین اسناد مشابه تشخیص داده شده مقایسه‌ای بین پاراگراف‌ها صورت می‌گیرد و از میان پاراگراف‌های موجود در این مرحله، پاراگراف‌هایی که با سند پرس‌وجو شباهت دارند به مرحله مقایسه در سطح جمله وارد می‌شوند. با این شیوه مقایسه چند سطحی تعداد مقایسات بین جملات اسناد بسیار محدود شده و ما می‌توانیم پردازش‌های بیشتری را در سطح جمله انجام دهیم. نتایج رتبه‌بندی نهایی اسناد نشان می‌دهد که سیستم پیشنهادی در کشف کپی‌برداری‌های دقیق و کپی‌برداری با تغییرات و همچنین برخی از کپی‌برداری‌های هوشمندانه مانند جایگزینی کلمات با مترادف‌هایشان موفق عمل می‌کند. از جمله کارهای آتی برای این پژوهش، بهبود روش خوشه‌بندی در سیستم پیشنهادی است. از آنجا که ارائه بهترین روش خوشه‌بندی هدف این مقاله نبوده است ما با استفاده از الگوریتم K-means سعی کرده‌ایم ایده بازیابی اسناد بر اساس خوشه‌بندی را در یک سیستم کشف دستبرد علمی آزمایش کنیم و بررسی و یافتن بهترین روش خوشه‌بندی در یک سیستم بازیابی اطلاعات و تأثیر آن بر عملکرد سیستم کشف دستبرد علمی، از اهداف بعدی ما در این زمینه خواهد بود؛ همچنین

توسعه سیستم پیشنهادی برای کشف انواع بیشتری از دستبردهای علمی هوشمندانه مانند خلاصه‌سازی متن نیز از اهداف دیگر ما خواهد بود.

مراجع

[1] کامران، ک.، احمدی، ع. و محسن زاده، م.، **"کشف سرقت ادبی در متون فارسی به کمک الگوریتم SimHash "**، یازدهمین کنفرانس سیستم‌های هوشمند ایران، 1391.

[2] M. Alzahrani, N. Salim, and A. Abraham, "*Understanding plagiarism linguistic patterns, Textual features, and detection Methods*", IEEE Trans. SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, vol. 42, no. 2, 2012.

[3] E. Stamatatos, "*Plagiarism Detection Using Stopword n-grams*", Journal of The American Society For Information Science And Technology, vol. 62, no. 12, pp. 2512-2527, 2011.

[4] G. S. Manku , A. Jain and A. D. Sarma, "*Detecting NearDuplicates for Web Crawling*", Data mining, 2007.

[5] T. W. S. Chow and M. K. M. Rahman, "*Multilayer SOM with treestructured data for efficient document retrieval and plagiarism detection*", IEEE Trans. Neural Netw., vol. 20, no. 9, 2009.

[6] H. Zhang and T. W. S. Chow, "*A coarse-to-fine framework to efficiently thwart plagiarism*", Elzevier Trans.Pattern Recognition., vol. 44, no. 2, pp. 471-487, 2011.

[7] E. Tiakas, D. Rafailidis, A. Dimou and P. Daras, "*MSIDX: Multi-Sort Indexing for Efficient Content-based Image Search and Retrieval*", IEEE Trans. Multimedia., vol. 15, no. 6, pp. 1415-1430, 2013.

[8] Shamsfard, M., and Kiani, S., and Shahedi, Y., "*STEP-1:Standard Text Preparation for Persion Language*", CAASL3 Third Workshop on Computational Approaches to Arabic Script- Languages.

[9] MR.Davarpanah, M.sanji and M.Aramideh, "*Farsi Lexical Analaysis and StopWord List*" Library Hi Tech, vol. 27, pp 435–449, 2009.

[10] Shamsfard, M., and Hesabi, A., and Fadaei H., and Mansoory, N., and Famian, A., and Bagherbeigi, S.,and Fekri, E.,and Monshizadeh, M., and Assi, SM., " *Semi Automatic Development of FarsNet; The Persian WordNet*", Proceedings of 5th Global WordNet Conference, 2010.

[11] M. Makrehchi, M. Kamel, "*A fuzzy set approach to extracting keywords from abstracts*", North American Fuzzy Information Processing Society- NAFIPS 2003,Banf, Canada, 2004.

[12] C. Grozea, C. Gehl, and M. Popescu, "ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection," in *Proc. SEPLN*, Donostia, Spain, 2012, pp. 10–18.

[13] C. Basile, D. Benedetto, E. Caglioti, G. Cristadoro, and M. D. Esposti,"A plagiarism detection procedure in three steps: Selection, matches and "squares," in *Proc. SEPLN*, Donostia, Spain, pp. 19–23.

[14] M. Elhadi and A. Al-Tobi, "Duplicate detection in documents and webpages using improved longest common subsequence and documents syntactical structures," in *Proc. 4th Int. Conf. Comput. Sci. Converg. Inf.Technol.*, Seoul, Korea, Nov. 2009, pp. 679–684.

[15] B. Stein, N. Lipka, and P. Prettenhofer, "Intrinsic plagiarism analysis," in *Language Resources & Evaluation*, 2010.

[16] M. zu Eissen, B. Stein, and M. Kulig, "Plagiarism detection without reference collections," in *Advances in Data Analysis*, 2007, pp. 359–366.

[17] S. zu Eissen and B. Stein, "Intrinsic plagiarism detection," in *Advances in Information Retrieval*, 2006, pp. 565–569.

[18] Z. Ceska, M. Toman, and K. Jezek, "Multilingual plagiarism detection," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence Lecture Notes in Bioinformatics)*, vol. 5253 LNAI, pp. 83–92, 2008.

[19] M. Potthast, A. Barr´on-Cede˜no, B. Stein, and P. Rosso, "Cross-language plagiarism detection," *Language Resources & Evaluation*, pp. 1–18, 2010.

[20] D. Pinto, J. Civera, A. Barr´on-Cede˜no, A. Juan, and P. Rosso, "A statistical approach to crosslingual natural language tasks," *J. Algorithms*, vol. 64, pp. 51–60, 2009.

[21] M. Kuta, J. Kitowski, "Optimisation of Character n-gram Profiles Method for Intrinsic Plagiarism Detection," *Springer Lecture Notes in Computer Science*, vol. 8468, pp 500-511, 2014.

[22] G. Oberreuter, G. L'Huillier, S. A. Ríos, J. D. Velásquez, "Outlier-Based Approaches for Intrinsic and External Plagiarism Detection," *Springer Lecture Notes in Computer Science*,vol 6882, pp 11-20, 2011.

[23] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey, "Scatter/Gather: A cluster-based approach to browsing large document collections," *In Proc. of SIGIR*, pages 318–329, 1992.

[24] A. Tombros, R. Villa, and C. van Rijsbergen, "The effectiveness of query-specific hierarchic clustering in information retrieval," *Inf. Process. Manage*, 38(4):559–582, 2002.

[25] A. Leuski, "Evaluating document clustering for interactive information retrieval," *In Proc. of CIKM*, pages 33–40, 2001.

[26] X. Liu and W. B. Croft, "Cluster-based retrieval using language models," *In Proc. of SIGIR*, pages 186–193, 2004.

زیرنویس‌ها

[1] Plagiarism
[2] Fingerprint
[3] Multi Layer Self-Organizing Map
[4] Multi-Level Overlapping
[5] Partitioning