استفاده از تکنیک مبتنی بر خوشه K-میانگین در تشخیص سرقت ادبی خارجی

چکیده

گروه بندی اسناد متنی یکی از عرصه‌های پژوهشی در حال ظهور است، که در آن اسناد شناخته شده هستند، به صورت دستی یا لگاریتمی دسته بندی شده و یا متمایز شده اند. مقاله بر استفاده از گروه بندی خودکار اسناد متنی در دامنه تشخیص سرقت ادبی تمرکز می‌کند. در جهان امروزی سرقت ادبی، به خصوص در زمینه آموزشی و پژوهشی یک نگرانی اصلی است. هدف این مقاله مطالعه و مقایسه روش‌های متفاوت گروه بندی اسناد در تشخیص سرقت ادبی خارجی است. در اینجا کانون اولیه کشف گروه بندی اسناد نظارت نشده / روش‌های خوشه بندی با استفاده از تغییرات متفاوت الگوریتم K-میانگین و مقایسه آن با روش مبتنی بر N-gram عمومی و روش مبتنی بر مدل فضای بردار است. سرانجام تحلیل و ارزیابی با استفاده از مجموعه داده ای از PAN-2013 ارزیابی شده است و عملکرد بر اساس precision، recall و efficiency از نظر زمان اجرای الگوریتم مقایسه می‌شود.

کلمات کلیدی: دسته بندی اسناد متنی، سرقت ادبی خارجی، بازیابی کاندید؛ N-gram؛ مدل فضای برداری، خوشه بندی K میانگین

1. مقدمه

دسته بندی خودکار اسناد متنی یا گروه بندی آن‌ها فرآیند اختصاص اسناد به یک گروه خاص یا بیشتر به صورت الگوریتمی است. هیچ سوالی مربوط به ارزش تجاری دسته بندی اسناد با توجه به محتوا وجود ندارد چرا که به خودی خود دارای هزاران کاربرد بالقوه است. با رشد سریع تکنولوژی، سربار اطلاعاتی به یک مشکل اصلی تبدیل شده است و

مرتب کردن و دسته بندی دقیق اسناد مربوطه به یک چالش تبدیل شده است. دسته بندی اسناد می‌تواند به صورت نظارت شده، نظارت نشده و نیمه نظارت شده گروه بندی شود. در دسته بندی سند نظارت شده، برخی از بازخوردهای خارجی برای ارائه اطلاعات دسته بندی صحیح در دسترس است. دسته بندی نظارت نشده، به عنوان خوشه بندی سند که دسته بندی را بدون هر ارجاعی به اطلاعات خارجی اجرا می‌کند اشاره دارد در حالی که بخش اصلی نیمه نظارت شده اسناد با استفاده از اطلاعات خارجی برچسب می‌خورند.

عرصه کاربردی دسته بندی اسناد، مانند شناسایی زبان، فیلتر کردن اسپم، دسته بندی ژانر، تحلیل احساسات، بازیابی اطلاعات (IR) و غیره بی شمار هستند. روش‌های دسته بندی اسناد در این دامنه کمتر کشف می‌شوند. رشد سریع فناوری شبکه، از جمله تعداد زیادی موتور جستجو، مخازن سند، سیستم‌های نرم افزاری ترجمه و غیره، نه تنها کانال‌های کسب دانش گوناگونی را ارائه می‌دهند، بلکه دری را به سوی سرقت ادبی متون باز می‌کنند. سرقت ادبی معمولا به استفاده نامشروع از اطلاعات، متن، ایده شخص دیگر، بدون ارجاع مناسب به منبع اصلی داده، اشاره دارد.

نوع و درجه سرقت ادبی از ساده ترین تا چالش برانگیزترین و پیچیده ترین محدوده بندی می‌شود. External/extrinsic و Internal/intrinsic دو روش تشخیص سرقت ادبی اصلی هستند. سرقت ادبی خارجی، سرقت ادبی را در تطبیق با یک سند اصلی یا بیشتر ارزیابی می‌کند در حالی که روش داخلی تغییر در سبک نوشتن منحصر یک نوشته را به عنوان شاخص سرقت ادبی بالقوه تحلیل می‌کند. این مقاله تشخیص سرقت ادبی خارجی را در جایی که اسناد منبع در فرم مجموعه داده/ وب منابع آنلاین در دسترس هستند توجیه می‌کند. در تشخیص سرقت ادبی خارجی، پنج وظیفه اصلی به صورت عمومی اجرا شدند. پیش پردازش مرحله اولیه ای است که اسناد منبع و مشکوک در معرض اصلاحات خاصی مانند حذف حرف‌های اضافه پرکابرد، tokenization، lowercasing، تقسیم بندی جمله، حذف نقطه گذاری قرار می‌گیرند. این به کاهش اندازه داده واقعی با حذف اطلاعات مربوطه با توجه به رویکرد استفاده شده کمک می‌کند. مرحله بعدی بازیابی کاندید در مواردی است که زیر مجموعه‌ای از اسناد مبنع بازیابی شده‌اند که به صورت کلی با اسناد مشکوک مشابه هستند. در مرحله سوم، مقایسه دقیق اسناد مشکوک در برابر مجموعه اسناد کاندید است که برای بازیابی بخش‌های سرقت ادبی شده انجام می‌شود. در اینجا تکنیک پردازش زبان

طبیعی (NLP) مانند قطعه قطعه کردن؛ تجزیه، برچسب گذاری بخشی از گفتگو (POS) می‌تواند پذیرفته شود. مرحله چهارم، تشخیص مرزهای عبور با تحمین دقیق مرزهای متون مورد سرقا واقع شده در اسناد منبع و اسناد مشکوک با استفاده از معیار اصلی است. سرانجام ارزیابی سیستم با استفاده از برخی از معیارهای استاندارد، یا معیارهای PAN برای رتبه بندی عملکرد آن صورت می‌گیرد.

دسته بندی اسناد متنی می‌تواند در فاز اولیه روش تشخیص سرقت ادبی خارجی برای بازیابی اسنادی که مشابه با اسناد مشکوک هستند استفاده شود، مانند بازیابی کاندید. این تعداد مقایسه اسناد را در مرحله تحلیل دقیق و بنابراین پیچیدگی را کاهش می‌دهد. مقاله بر مطالعه جامع و تحلیل روش‌های مبتنی بر N-gram، روش‌های مبتنی بر مدل فضای برداری (VSM) و روش‌های مبتنی بر خوشه بندی K میانگین در زمان استفاده از مراحل بازیابی کاندید تشخیص سرقت ادبی خارجی تمرکز می‌کند. روش‌ها بر اساس تکنیک‌های NLP مانند ریشه یابی، ریشه یابی و قطعه بندی مقایسه می‌شوند.

2. کارهای مربوطه

در تشخیص سرقت ادبی، بازیابی کاندید می‌تواند به عنوان کار بازیابی اطلاعات دیده شود. بر اساس این مفهوم، روش‌های بازیابی کاندید متفاوت با استفاده از مدل‌های IR توسط محققین برجسته بیان شده است. در این گروه، روش‌های مبتنی بر اثر انگشت و N-gram، روش‌های مبتنی بر هش و مدل‌های فضای بردار محبوبیتی کسب کردند. S Schleimer یک روش مبتنی بر اثر انگشت را پیشنهاد داد که در آن سند به K-grams تقسیم می‌شود که در آن "K" یک مقدار مشخص شده توسط کاربر است. سپس هر K-gram هش می‌شود و برخی از آن‌ها به عنوان اثر انگشت سند با استفاده از اندازه خاص پنجره انتخاب می‌شوند. نتایج با استفاده از Stanford Webbase ارزیابی شده اند. روش فقط می‌تواند تطبیق طول را تشخیص دهد.William B. Cavnar & John M. Trenkle یک روش گروه بندی متن مبتنی بر N-gram را ارائه می‌دهد که در ان اسناد اولیه به عنوان ویژگی N-gram ارائه شده است. سپس با استفاده از برخی از معیارهای فاصله تشابه بین اسناد محاسبه می‌شود و آن‌ها دسته بندی می‌شوند. این

فرکانس N-gram را با استفاده از قانون 'Zipf برای در نظر گرفتن N-grams برتر محاسبه می‌کند. برای تست، داده از پنج گروه تازه استفاده می‌شود و نتایج خوبی برای دسته بندی انواع مقالات جدید ارائه می‌شود. Peter Nather یک روش مشابه را پیشنهاد داد که در آن 1-gram تا 5-gram برای آزمایشات استفاده شدند. معیار تشابه دو پروفایل N-gram در مجموع تفاوت بین رتبه N-gramها در یک پروفایل و رتبه در دیگر پروفایل اتخاذ شده است. روش با استفاده از داده ای از Project Gutenberg ارزیابی شده است و دریافتیم که روش به خوبی با اسناد کوتاه اجرا شده است. با حجم بزرگ و دستکاری‌ها د محتوا عملکرد روش کاهش می‌یابد.

در [8]، 5 تا 6 N-grams برای نمایش پروفایل‌های اسناد تشکیل شده اند. سپس شناسه سند به یک مجموعه از ارزش‌های هش ساختار یافته نگاشت می‌یابد. یک شاخص معکوس شده، مقدار هش را به دنباله ای از شناسه سند نگاشت می‌کند که از آن محاسبه شده است. تشابه بیشتر با استفاده از ضریب جاکارد محاسبه می‌شود و سیستم با استفاده از مجموعه داده PAN-09 ارزیابی می‌شود. یک رویکرد بازیابی کاندید مبتنی بر N-gram در [9] پیشنهاد شده است که از رویکرد ضریب جاکارد و شینگل در مرحله بازیابی کاندید استفاده می‌کند. در اینجا بعد از اینکه پیش پردازش پروفایل‌های N-gram با N= 3 یا 4 صورت گرفت سپس تشابه جاکارد محاسبه می‌شود. یک آستانه 0.1 برای بازیابی مجموعه اسناد کاندید استفاده می‌شود. ارزیابی سیستم نهایی با استفاده از مجموعه داده PAN-10 انجام شده است. Rajiv Yerra و Yiu --Kai Ng [10] یک روش تشخیص کپی مبتنی بر جمله را برای اسناد وب اجرا کردند. در اینجا تشابه بین جملات سند با استفاده از سه رویکرد 4-gram کمتر تکراری محاسبه شده است، در جایی که نشانه‌های اصلی در نظر گرفته شده اند. رویکرد IR فازی نیز بحث شده است، که بهتر از مورد اول اجرا می‌شود.

مدل یازیابی مشهور دیگری مدل فضای برداری (VSM) است که در وظیفه بازیابی کاندید استفاده شد. در [11] 16 grams ایجاد می‌شوند که سپس به عنوان بردارهای استفاده کنند از مقادیر tf-idf نمایش داده می‌شوند و فاصله سند بیشتر با استفاده از شباهت کسینوسی اندازه گرفته می‌شود. Muhr و همکاران [12] یک رویکرد مبتنی بر VSM را پیشنهاد کردند، در جایی که برداربندی اولیه اسناد منبع و مشکوک انجام شده است. سپس نزدیک ترین سندی که با سند مشکوک متناظر است از مجموعه مرجع با استفاده از شباهت کسینوسی یافت می‌شود. رویکرد تشابه

در [13] و [14] برای بازیابی کاندید بحث شده است و سیستم با استفاده از مجموعه داده PAN ارزیابی شده است. در روش‌های VSM زمان اجرا به سرعت با مجموعه داده بزرگ افزایش می‌یابد.

نگاشت‌های خود سازمان ده (SOM) در [15] برای سازمان دهی و خوشه بندی داده استفاده شده است. این مقاله از یک SOM ساختار یافته با درخت چند لایه (MLSOM) استفاده می‌کند که در ان داده گره در سطوح متفاوت یک درخت در لایه‌های متفاوت MLSOM پردازش شده است. Liping Jing و همکاران [16] یک رویکرد خوشه بندی سند که از الگوریتم k-میانگین برای خوشه بندی داده پراکنده استفاده می‌کند، ارائه داده‌اند. این به صورت خودکار وزن کلمات کلیدی را در هر خوشه برای شناسایی اهمیت آن‌ها محاسبه می‌کند. آزمایشات با استفاده از داده‌هایی از 20 گروه خبری انجام شده است و نتایج امیدوار کننده است. کارهای متفاوت با استفاده از تکنیک خوشه بندی در IR و دامنه دسته بندی انجام شده است. اما زمانی که این دامنه تشخیص سرقت ادبی خارجی بوجود می آید، استفاده از تکنیک خوشه بندی بسیار محدود می‌شود.Duo Zuo و همکاران [17] یک روش تشخیص سرقت ادبی خارجی را پیشنهاد دادند که از خوشه بندی برای پردازش پست استفاده می‌کند. در اینجا در درجه اول مرحله پیش انتخاب برای کاستن از دامنه تشخیص با استفاده از اثرانگشت‌های پی در پی انجام شده است. سپس همه بخش‌هایی بین دو سند یافت می‌شوند و ادغام می‌شوند. خوشه بندی بیشتر برای کاهش تاثیر متن مبهم استفاده می‌شود. سیستم در نهایت با استفاده از مجموعه داده PAN-10 ارزیابی شده است. بنابراین از بررسی ادبیات موضوعی انجام شده، مشاهده شده است که مدل‌های IR، با روش‌های مبتنی بر N-gram و VSM به صورت گسترده در تشخیص سرقت ادبی خارجی استفاده می‌شود در حالی که تکنیک خوشه بندی کمتر استفاده می‌شود. روش دسته بندی سند نظارت نشده و روش خوشه بندی سند پتانسیل خوبی در مرحله بازیابی کاندید تشخیص سرقت ادبی خارجی دارد. در این مقاله، خوشه بندی سند با استفاده از الگوریتم K میانگین برای وظیفه بازیابی کاندید کشف شده است. بسط متفاوتی از K-میانگین پایه با استفاده از تکنیک NLP و N-grams نیز بحث شده است. الگوریتم‌ها با استفاده از مجموعه داده جزئی از مجموعه PAN-13 و مقایسه با روش‌های مبتنی بر N-gram پایه و روش مبتنی بر VSM کلاسیک ارزیابی شدند. روش‌های پیشنهاد شده و مقایسه شده در بخش‌های زیر بحث شدند.

3. روش‌های پیشنهاد شده و مقایسه شده

این بخش روش‌های گوناگون پیشنهاد شده و مقایسه شده را در این مقاله تشریح می‌کند. روش‌های زیر بحث شده اند:

1. روش مبتنی بر N-gram
2. روش مدل فضای برداری (VSM)
3. روش مبتنی بر خوشه با استفاده از الگوریتم K-میانگین
4. K-میانگین با stemming
5. K-میانگین با ریشه یابی
6. K-میانگین با N-grams
7. K-میانگین با قطعه بندی

در آغاز برخی از پیش پردازش‌های سند متنی انجام می‌شوند. این پیش پردازش‌ها شامل tokenization، حذف نقطه گذاری و حذف کلمات اضافی (کلماتی بدون معنا) است. یک لیست از 50 کلمه بیشتر استفاده شده در زبان انگلیسی توسط British National Corpus ارائه شده است که شامل 90 میلیون نشانه است که معمولا استفاده می‌شوند. در روشی که از قطعه بندی استفاده می‌شود، ابتدا چانک‌ها تشکیل می‌شوند و سپس کلمات اضافی حذف می‌شود. ابتدا روش مبتنی بر N-gram سنتی و روش VSM سنتی برای بازیابی کاندید استفاده می‌شود. روش مبتنی بر خوشه با استفاده از الگوریتم K-میانگین پیشنهادی پذیرفته شده است. تغییرات متفاوت تر K-میانگین صورت می‌گیرند و نتایج تحلیل و مقایسه می‌شود.

3.1 روش مبتنی بر N-gram

در اینجا پیش پردازش عمومی انجام می‌شود. سپس سند به N-gramها و N-shingles تقسیم می‌شود. این به دنباله ای از کلمات پی در پی با اندازه ”N“ اشاره دارد، در اینجا ”N“ کاربر مشخص شده است. هر دو سند مشکوک و سند منبع به پروفایل‌های N-gram خود تبدیل می‌شوند و تشابه با استفاده از ضریب Dice محاسبه می‌شود. این

مشابه با ضریب جاکارد است اما اثر عبارات مشترک بین اسناد را کاهش می‌دهد. فرض کنید X_{suspng} و X_{srcng} مشخصه‌های N-gram مشکوک و منبع باشند، سپس ضریب Dice به شکل زیر تعریف می‌شود:

$$Dice(X_{suspng}, X_{srcng}) = 2\frac{\left|X_{suspng} \cap X_{srcng}\right|}{\left|X_{suspng}\right| + \left|X_{srcng}\right|} \qquad (1)$$

جمله انگلیسی زیر را در نظر بگیرید (E1):

“The people left their countries and sailed with Gilbert.”

بعد از پیش پردازش ابتدای، نشانه‌هایی بدست می‌آید (E1-tokens):

['people', 'left', 'countries', 'sailed', 'Gilbert']

پس از تشکیل trigram ؛ N=3 ؛ (E1-3-gram) داریم:

']]people', 'left', 'countries'], ['left', 'countries', 'sailed'], ['countries', 'sailed', 'Gilbert'], ['sailed', 'Gilbert[['

بعد از تشکیل مشخصه‌های N-gram منبع و مشکوک، تشابه با استفاده از (1) محاسبه می‌شود. در عوض استفاده از مقدار آستانه برای انتخاب مجموعه کاندید صورت می‌گیرد، در اینجا تشابه بین هر منبع با همه موارد مشکوک محاسبه شده است. سپس اسناد مشکوک دارای حداکثر معیار ضریب جاکارد انتخاب شده به عنوان سند مربوطه است.

3.2 روش مدل فضای بردار (VSM)

مدل فضای برداری (VSM) یک مدل جبری است که نشان دهنده اطلاعات متنی به عنوان یک بردار است. در اینجا، بعد از پیش پردازش اولیه مورد نیاز، یک دیکشنری از عبارات (کلمات) از هر سند منبع استخراج می‌شود که با همه اسناد مشکوک مقایسه می‌شود. VSM اهمیت کلماتی که به صورت مکرر استفاده شده اند با متریک فراوانی سندی معکوس (tf-idf) نشان می‌دهد. فراوانی سندی معکوس idf(t) سپس محاسبه می‌شود که تاکید دارد که یک عبارت که تقریبا در کل مجموعه اسناد موجود است خوب نیست. در آخر، tf-idf محاسبه می‌شود و تشابه بین بردارهای سند

با استفاده از شباهت کسینوسی محاسبه می‌شود. شباهت کسینوسی بین دو سند X_{susp} (مشکوک) و X_{src} (منبع) به شکل زیر محاسبه می‌شود:

$$Cos(X_{susp}, X_{src}) = \frac{V(X_{susp}) \bullet V(X_{src})}{\|V(X_{susp})\| \, \|V(X_{src})\|} \qquad (2)$$

در اینجا $V(X_{susp})$ و $V(X_{src})$ نمایش بردار سند مشکوک و منبع را به ترتیب نشان می‌دهند. صورت کسر در معادله (2) به ضرب نقطه ای سه بردار و مخرج کسر به ضرب نرم اقلیدسی آن‌ها اشاره دارد. بعد از محاسبه تشابه با استفاده از معادله (2)، اسناد کاندید توسط رویکرد مشابه با روش N-gram انتخاب می‌شود. بنابراین هر سند منبع با همه سندهای مشکوک مقایسه می‌شود و سند مشکوک با حداکثر شباهت کسینوسی انتخاب می‌شود. باید اشاره کنیم که مواردی وجود دارند که در آن یک سند منبع به هر سند مشکوکی بی ربط است، مانند زمانی که منبع کل وب است. اما رویکرد غیر آستانه ای در بازیابی کاندید کارامد است چرا که اینجا تحلیل دقیق تری برای تشخیص اسنادی که به واقع سرقت ادبی شده اند صورت می‌گیرد.

3.3 روش مبتنی بر خوشه با روش پیشنهادی الگوریتم K-میانگین

در این روش رویکرد خوشه بندی استفاده شده است، که در آن اسناد مشابه همراه هم به عنوان یک خوشه گروه بندی می‌شوند. در اینجا الگوریتم استفاده شده K-میانگین است که یک تکنیک خوشه بندی بخشی کارامد است. در الگوریتم خوشه بندی K-میانگین دو پارامتر اصلی تعداد خوشه‌ها (K) و مرکز خوشه اولیه/مرکزیت است. الگوریتم پایه به شرح زیر است:

1. انتخاب k و مرکزیت اولیه، در اینجا "K" تعداد مرکزیتی است که باید انتخاب شود.
2. انتساب هر شی به گروهی که با استفاده از معیار فاصله یا تشابه کمترین مرکزیت را دارد.
3. زمانی که همه شی‌ها تخصیص داده شدند، موقعیت‌های مرکزهای "K" دوباره محاسبه شود.
4. تکرار مرحله 2 و 3 تا زمانی که مرکزیت دیگر تغییر نکند.

در اینج مسئله اصلی تصمیم گیری بر مقدار 'K' و مرکزیت اولیه است، چرا که این دو پارامترها کاملا نتایج الگوریتم را تنظیم می‌کنند. با در نظر گرفتن این محدودیت، رویکرد پیشنهادی این پارامترها مقادیر ثابتی را ارائه می‌دهند. در اینجا مفهوم استفاده شده این است که هر سند مشکوک به عنوان یک مرکزیت عمل می‌کند. سند منبع که به صورت کلی مشابه با سند مشکوک است برای خوشه‌هایی که شامل این سند مشکوک به عنوان مرکزیت هستند؛ گروه بندی می‌شود. بنابراین 'K' به عنوان مجموع تعداد اسناد مشکوک در نظر گرفته می‌شود، فرض کنید که مجموعه مشکوک (مجموعه PAN) داده شده باشد. با استفاده از این مفهوم، الگوریتم K-میانگین پایه برای وظیفه بازیابی کاندید به صورت زیر اصلاح می‌شود:

1. K= تعداد اسناد مشکوک
2. تنظیم مرکزیت K اولیه = هریک از K سند مشکوک.
3. تخصیص هر سند منبع به خوشه با نزدیک ترین مرکزیت با استفاده از شباهت کسینوسی

فرآیند با استفاده از یک پکیج Python اجرا می‌شود که خوشه بندی را تسهیل می‌کند و زمان در نظر گرفته شدن را برای بردارهای سند کاهش می‌دهد و تشابه آن‌ها را محاسبه می‌کند. سپس بر اساس معیار تشابه اسناد منبع آن‌ها بر اساس تناظر اسناد مشکوک گروه بندی می‌شوند. بنابراین هر خوشه متناظر با مجموعه کاندید از اسناد برای یک سند مشکوک خاص است. در الگوریتم پیشنهاد شده، زمانی که مرکزیت‌ها تثبیت شوند، تنها یک تکرار نیاز است. این از نظر پیچیدگی زمانی کارامد است.

3.4 تغییرات الگوریتم K-میانگین پیشنهادی

روش بازیابی کاندید K-میانگین بحث شده در زیر بخش 3.2، به عنوان رویکرد پایه استفاده شده است. سپس بسط مرکزیت بوجود می‌آیند و الگوریتم ارزیابی می‌شود. این روش‌ها در بخش‌های زیر بحث شده اند.

3.4.1 K-میانگین با Stemming (K-Stem)

در این روش، تنها تغییر این است که بعد از نشانه گذاری سند، stemming صورت می‌گیرد. ریشه یابی یک فرآیند اکتشافی از حذف وندها از کلمات است. مراحل باقی مانده مشابه با رویکرد پایه انجام شده است.

3.4.2 K-میانگین با ریشه یابی (K-Lem)

این روش از ریشه یابی به جای Stemming استفاده می‌کند. محدودیت فرم‌های پایه دیکشنری یک کلمه و موفولوژی را مورد استفاده قرار می‌دهد . این رابطه تنگاتنگی با ریشه یابی دارد اما ریشه یابی تنها در یک کلمه تکی اعمال می‌شود در حالی که محدودیت سازی بر کل متن اعمال می‌شود . این می‌تواند تبعیض قائل شدن بین کلماتی باشد که بسته به بخشی از گفتار دارای معانی متفاوت هستند.

جدول ۱:آمارهای داده

No. of Documents		
	سند مشکوک	سند متبع
Set 1	39	205
Set 2	31	213
Set 3	35	209

3.4.3 K-میانگین با N-gram (K-Ng)

در اینجا روش K-میانگین با روش مبتنی بر N-gram ترکیب شده است. به جای اتخاذ کلمات منحصر N-grams ایجاد شده و پیش پردازش بیشتر از رویکرد پایه استفاده می‌شود.

3.4.4 K-میانگین با قطعه بندی (K-Chk)

روش از قطعه بندی برای تشکیل عبارات گرامری به جای مقابله با unigrams استفاده می‌کند. در ابتدا یک درخت تجزیه ساخته می‌شود. سپس اسم، فعل، صفت، قید از آن استخراج می‌شود چرا که این عبارات در معنا سازی یک جمله نقش دارد. الگوریتم با استفاده از مراحل پیش پردازش ارزیابی می‌شود، ریشه یابی و محدودیت در زیر بخش‌های 3.3.1 و 3.3.2 بحث شده اند؛ مانند K-Chk-Stem و K-Chk-Lem.

4. تنظیمات آزمایشی و تحلیل نتایج

4.1آماره‌های داده

الگوریتم‌ها با استفاده از سه مجموعه از اسناد از مجموعه PAN-13 ارزیابی شدند. هر مجموعه دارای اسناد مشکوک و اسناد منبع متناظر هستند که در جدول 1 آورده شده است. سه مجموعه داده استفاده شده است:

- Set-1: بدون ابهام
- Set-2: ابهام تصادفی
- Set-3: ابهام در ترجمه

هیچ مجموعه مبهمی شامل جفت سند نیست و در ان اسناد مشکوک کپی‌های دقیقی از سند منبع هستند. در ابهام تصادفی عبارت در جفت سند عبارت با بهم آمیختن، جایگزینی و غیره مبهم می‌شود. مجموعه ابهام ترجمه شامل جفت اسنادی است که عبارات آن از طریق توالی از تفسیرها اجرا می‌شود. از اینجا خروجی یک ترجمه به ورودی ترجمه بعد منتقل می‌شود و زبان ترجمه نهایی مشابه با زبان سند اصلی است. با توجه به محدودیت‌های سخت افزاری تنها برخی از اسناد از این مجموعه برای ارزیابی انتخاب می‌شوند.

4.2 ارزیابی

روش‌های بحث شده در بخش 3 با استفاده از معیارهای، Recall (rec)، precision (prec) و زمان اجرا ارزیابی می‌شود. در سناریو بازیابی کاندید، recall با تعداد اسناد بازیابی مربوطه با تعداد واقعی اسناد مربوطه بازیابی شده تعریف می‌شود.

$$rec = \frac{\#\ of\ \ relevant\ \ documents\ \ retrieved}{Actual\ \ \#\ of\ \ relevant\ \ documents} \qquad (3)$$

precision به عنوان تعداد اسناد مربوطه بازیابی شده با مجموع تعداد اسناد بازیابی شده توسط سیستم تعریف شده است.

$$prec = \frac{\#\ of\ \ relevant\ \ documents\ \ retrieved}{Total\ \ \#\ of\ \ documents\ \ retrieved\ \ by\ system} \qquad (4)$$

4.3 نتایج و بحث‌ها

هر مجموعه با استفاده از الگوریتم‌های بحث شده در بخش 3 ارزیابی شده است و نتایج مقایسه شده اند. آماره‌های داده ارائه شده در جدول 1 برای ارزیابی استفاده شده است. نتایج بدست آمده برای هر مجموعه از نظر precision، recall و زمان اجرا با استفاده از روش‌های بحث شده در بخش 3 در شکل 1، شکل 2 و شکل 3 داده می‌شود. در اینجا precision و recall در محور اولیه و زمان اجرا در محور ثانویه نشان داده می‌شود. با تحلیل شکل 1، می‌توان مشاهده کرد که روش‌های مبتنی بر N-gram نتایج خوبی برای Set-1 (مجموعه بدون ابهام) می‌دهند. این مشاهده کرده است که با افزایش مقدار "N"، هردو recall و precision کمی کاهش می‌یابند. در این مقاله، تنها روش VSM کلاسیک پایه اجرا می‌شود و ارزیابی می‌شود. با تحلیل هر یک از این نمودارها، می‌توان متوجه شد که روش VSM یک precision به شدت کاهش یافته و recall را در هر سه مجموعه ارائه می‌دهد. این می‌تواند مشاهده کند که این روش در مقایسه با رویکردهای دیگر نیازمند زمان اجرای بالایی است. روش VSM در زمانی که همراه با رویکردهای مانند رتبه بندی سند، فهرست بندی معنایی نهان (LSI) استفاده می‌شود، نتایج خوبی ارائه می‌دهد. اما

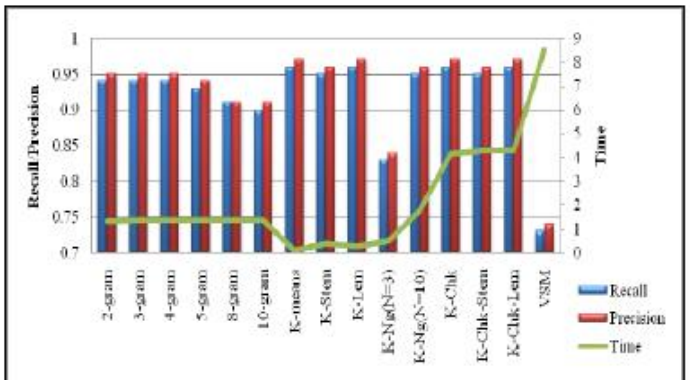
پیچیدگی زمانی هنوز محدودیت اصلی این روش است به خصوص زمانی که با مجموعه داده‌های بزرگی سر و کار دارد. از نمودار در شکل 1، دیده می‌شود که برای روش K-میانگین پیشنهادی، precision و recall در مقایسه با روش N-gram و VSM افزایش می‌یابد. در روش K-میانگین زمان اجرا به صورت قابل توجهی کاهش می‌یابد. روش‌های K-Chk و K-Ng (N=10) نیز precision و recall خوبی ارائه می‌دهند اما زمان اجرا افزایش می‌یابد. این به این دلیل است که تشکیل چانک‌ها در روش K-Chk و N-grams در روش K-Ng زمان بر است. بنابراین روش K-میانگین از دو رویکرد دیگر در زمانی که اسناد بدون ابهام (Set-1) هستند بهتر اجرا می‌شود. همه نتایج حداکثر 3 تا 5 نقطه متفاوت را نشان می‌دهند. این به این دلیل است که تنها مجموعه داده کوچکی برای ارزیابی در نظرگرفته می‌شود.

شکل 2 روش‌های استفاده از Set-2 را تحلیل و ارزیابی می‌کند، برای مثال ابهام تصادفی. می‌توان مشاهده کرد که روش K-میانگین و تغییرات آن، با K-Chk و K-Ng (N=10) بهتر از روش های مبتنی بر N-gram و VSM اجرا می‌شود. در روش مبتنی بر N-gram N=3 و N=4 نتایج خوبی ارائه می‌دهند. همچنین دریافته شده است که روش K-میانگین زمان اجرای بسیار کمتری برای Set-2 نیاز دارد. نمودارها ارزیابی روش با استفاده از Set-3 رسم شده‌اند. پیچیدگی ابهام در این مجموعه در مقایسه با دو مجموعه دیگر بالا است. با تحلیل گراف در شکل 3 و مقایسه روش‌های متفاوت، می‌توان اشاره کرد که روش خوشه بندی K-میانگین از روش N-gram و روش‌های مبتنی بر N-gram در زمان ارزیابی با استفاده از Set- 3 برتر است. از نظر precision و recall میانگین-k و K-Ng (N=10) از دو روش دیگر بهتر اجرا می‌شود اما KNg (N=10) به زمان اجرای به طور قابل توجه زیادی نیاز دارد.

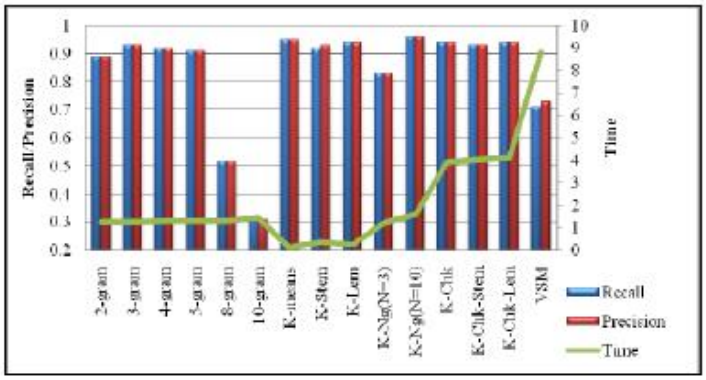
از مقایسه‌ها و تحلیل‌های صورت گرفته، مشاهده شده است که روش K-میانگین در زمان سر و کار داشتن با داده‌های به شدت مبهم در مقایسه با دو رویکرد دیگر نتایج امیدوار کننده ای ارائه می‌دهند. می‌توان نتیجه گرفت که از نظر زمان اجرا الگوریتم K-میانگین پیشنهادی کارآمدی اجرا را با همه مجموعه‌های بحث شده در بخش 4.1 ارئه می‌دهد. بنابراین روش پیشنهادی مراحل بازیابی کاندید را بدون به خطر انداختن precision بازیابی تسریع میکند. هر efficiency الگوریتم به precision و زمان بستگی دارد، از این رو توازن بین این دو فاکتور در نظر گرفته می‌شود.

5. نتیجه گیری و تحقیقات آینده

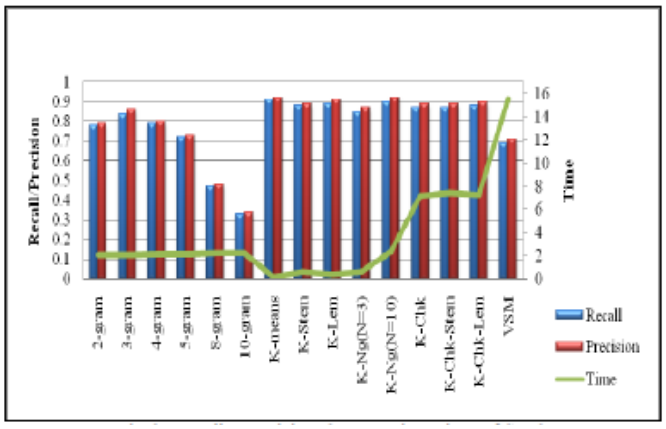
مقاله، کاربرد گروه بندی خودکار اسناد متنی را در تشخیص سرقت ادبی خارجی بررسی کرده است. در اینجا تلاشی برای استفاده از تکنیک خوشه بندی در مرحله بازیابی کاندید تشخیص سرقت ادبی خارجی صورت گرفته است.



شکل 1: recall، precision و زمان اجرای Set1



شکل 2: recall، precision و زمان اجرای Set2



شکل 3: recall، precision و زمان اجرای Set3

بازیابی کاندید کارامد می‌تواند به کاهش تعداد مقایسه‌های اسناد کمک کند و از این رو پیچیدگی زمانی را در مرحله تحلیل محلی جامع تشخیص کاهش دهد. در اینجا روش‌های مبتنی بر N-gram، رویکرد VSM پایه، و روش‌های مبتنی بر خوشه K-میانگین تحلیل و مقایسه شدند. یک رویکرد K-میانگین جدید برای بازیابی اسناد کاندید پیشنهاد شد. مقاله بر تغییرات اصلی که می‌توانند برای بسط الگوریتم K-میانگین پیشنهادی استفاده شوند سایه انداخته است. از نتایج و بحث‌ها در بخش 4 می‌توان نتیجه گرفت که الگوریتم خوشه بندی میانگین-K از هر دو روش VSM و N-gram برتر است. روش‌ها نتایج امیدوار کننده ای برای اسنادی می‌دهند که با استفاده از روش‌هایی ماند ترجمه که به سختی تشخیص داده می‌شود، مبهم شده اند. تغییرات K-میانگین پیشنهادی از تکنیک‌های NLP متفاوت با K-Ng (N=10)، K-Stem، K-Lem، K-Chk، K-Chk-Stem، K-Chk-Lem استفاده می‌کند که recall و precision خوبی را حاصل می‌کند. از نظر زمان اجرا، که فاکتور مهمی در هر سیستم نرم افزاری است، روش پیشنهادی به صورت موثر اجرا می‌شود. این وظیفه بازیابی سند کاندید را با کاهش قابل توجه در زمان اجرا، سرعت می‌بخشد.

برای کارهای آینده، رویکرد K-میانگین با Word Net می‌تواند برای مقایسه بهتر اسناد، در شرایط آگاهانه تر استفاده شود. اسنادی که به صورت هوشمندانه‌ای با استفاده از مترادف‌ها دستکاری شده اند می‌توانند بازیابی شوند. برای بهبود recall، الگوریتم‌های K-میانگین-فازی می‌توانند استفاده شوند که خوشه بندی نرم را ارائه می‌دهند. تکنیک‌های چند پردازنده ای و چند نخی برای بهبود efficiency زمانی الگوریتم‌ها به خصوص در زمان سر و کار داشتن با مجموعه داده‌های بزرگ استفاده شدند.

6. REFERENCES

[1] Peter Jackson and Isabelle Moulinier,Natural Language Processing for Online Applications:Text Retrieval, Extraction , and Categorization,JAN 2002,pp.119-225.

[2] Arzucan Ozgur, "Supervised and Unsupervised Machine Learning Techniques for Text Document Categorization",MSc.Bogazici University,2004.

[3] Ahmed Hamza Osman1, Naomie Salim and Albaraa Abuobieda, "Survey of Text Plagiarism Detection", Journal of Computer Engineering and Applications vol.1, June 2012.

[4] Salha M. Alzahrani, NaomieSalim, and Ajith Abraham, " Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods", IEEE transactions on systems, man, and cybernetics,vol.42 no.2, march 2012.

[5] S. Schleimer, D. Wilkerson, and A. Aiken, "Winnowing: Loca algorithms for document fingerprinting," In *Proc. ACM SIGMOD Int Conf.Manage. Data*, New York, 2003, pp. 76–85.

[6] William B. Cavnar , and John M. Trenkle, "N-Gram-Based Tex Categorization", In Proc. of SDAIR-94, 3rd Annual Symposium On Document Analysis And Information Retrieval.

[7] Peter Nather,"N-Gram-Based Text Categorization",Thesis. Bratislva University,2005.

[8] J.Kasprzak,M.Brandejs, and M.K ripac, "Finding plagiarism by evaluating document similarities," In *Proc. SEPLN*, Donostia, Spain, pp 24–28.

[9] S. Alzahrani, "Plagiarism auto-detection in arabic scripts using statement-based fingerprints matching and fuzzy-set information retrieval approaches," M.Sc. thesis, Univ. Technol. Malaysia, Johor Bahru, 2008.

[10] R. Yerra and Y.-K. Ng, "A sentence-based copy detection approach for web documents," in *Fuzzy System and Knowledge Discovery*, 2005, pp. 557–570.

[11] 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.

[12] M. Zechner, M. Muhr, R. Kern, and M. Granitzer, "External and intrinsicplagiarism detection using vector space models," in *Proc. SEPLN*, Donostia, Spain, pp. 47–55.

[13] Asif Ekbal, Sriparna Saha and Gaurav Choudhary, " Plagiarism Detection in Text using Vector Space Model, In Proc. of 12th International Conference on Hybrid Intelligent Systems(HIS), pp.366-371, Pune, 2012.

[14] Rasia Naseem and Sheena Kurian, " Extrinsic Plagiarism Detection in Text Combining VSM and Fuzzy Semantic Similarity Scheme", Journal of Advanced Computing, Engineering and application(IJACEA),vol.2, December 2013.

[15] 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, pp. 1385–1402, Sep. 2009.

[16] D. Zou, W.Long, and Z. Ling, "A cluster-based plagiarism detection method - Lab report for PAN at CLEF 2010", In Proc. of the 4th Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse. Padua,Italy, 2010.

[17] Liping Jing, Michael K.Ng,Jun Xu, and Joshua Zhexue Huang, "Sub-space Clustering of text documentswith feature weighting K-means algorithm", In Proc. of the 9th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining,pp.802-812,Berlin,2005.

[18] Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification,2nd ed.,Wiley:2000.

[19] Martin Potthast, Matthias Hagen, Tim Gollub, Martin Tippmann(et al.),"Overview of 5th International Competition on Plgiarism Detection, CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September,Valencia, Spain. ISBN 978-88-904810-3-1. ISSN 2038-4963. 2013.

[20] Ch. Aswani Kumar, and S. Srinivas, "On the Performance of Latent Semantic Indexing-based Information Retrieval," Journal of Computing and Information Technology - CIT 17,2009, pp.259–264.