مقاله پژوهشی

ارائه یک موتور جستجو برای بازیابی رویداد ساختارمند از منابع خبری

علیرضا میرزائیان و صادق علی‌اکبری

چکیده: تحلیل محتوای اخبار منتشرشده، یکی از مسایل مهم در حوزه بازیابی اطلاعات است. امروزه تحقیقات زیادی برای تحلیل تک‌تک مقالات خبری انجام شده است، در حالی که اکثر رویدادهای خبری به شکل چندین مقاله مرتبط به هم به طور مکرر در رسانه‌ها منتشر می‌شوند. تشخیص رویداد، وظیفه کشف و گروه‌بندی اسنادی را دارد که رویدادی یکسان را شرح می‌دهد و با ارائه یک ساختار قابل درک از گزارش‌های خبری، هدایت بهتر کاربران در فضاهای خبری را تسهیل می‌کند. با رشد سریع و روزافزون اخبار برخط، نیاز به ایجاد موتورهای جستجو برای بازیابی رویدادهای خبری به منظور تسهیل جستجوی کاربران در این فضاهای خبری بیش از پیش احساس می‌شود. فرض اصلی تشخیص رویداد بر این است که به احتمال زیاد کلمات مرتبط به یک رویداد یکسان در دنیای واقعی، در اسناد و پنجره‌های زمانی مشابه ظاهر می‌شوند. بر همین اساس ما در این تحقیق روشی گذشته‌نگر و ویژگی‌محور پیشنهاد می‌کنیم که کلمات را بر اساس ویژگی‌های معنایی و زمانی گروه‌بندی می‌کند. سپس از این کلمات برای تولید یک بازه زمانی و توصیف متنی قابل درک برای انسان استفاده می‌کنیم. ارائه یک معماری مناسب و استفاده مؤثر از خوشه‌بندی جهت بازیابی رویدادها و همچنین تشخیص مناسب زمان رویداد، از نوآوری‌های این پژوهش به شمار می‌روند. روش پیشنهادی روی مجموعه داده AllTheNews که تقریباً شامل دویست هزار مقاله از ۱۵ منبع خبری در سال ۲۰۱۶ می‌باشد ارزیابی شده و با روش‌های دیگر مقایسه گردیده است. ارزیابی‌ها نشان می‌دهد که روش پیشنهادی در دو معیار دقت و یادآوری نسبت به روش‌های پیشین عملکرد بهتری دارد.

***کلیدواژه:* تشخیص رویداد، موتور جستجو، بازیابی اطلاعات، متن‌کاوی.**

۱- مقدمه

در سال‌های اخیر حجم اخبار منتشرشده در اینترنت به شدت افزایش یافته است. رسانه‌های خبری با گزارش آخرین رویدادهای سراسر جهان، نقش مهمی در زندگی روزمره مردم ایفا کرده‌اند. علاوه بر این با ظهور دستگاه‌های تلفن همراه با قابلیت اینترنت مثل GPRS و ۳G و ۴G، اینترنت به تدریج به مهم‌ترین منبع اطلاعاتی جهانی تبدیل شده است. تغییر رویه عظیم منابع خبری، روش دسترسی مردم به اخبار را تغییر داده است. برای مثال، متخصصان اقتصادی که به اطلاعات بلادرنگ برای تصمیم‌گیری نیاز دارند، در حال حاضر از سرویس‌های پیام‌رسان مثل تلگرام به عنوان یک رسانه تحویل ۲۴ساعته استفاده می‌کنند. همچنین ارائه‌دهندگان اخبار سنتی مثل BBC و CNN به اینترنت مهاجرت کرده‌اند و ویژگی‌های چندرسانه‌ای مانند تصویر و ویدئو از گزارش‌های خبری ارائه می‌دهند. بمباران همه‌جانبه اطلاعات خبری، با خطر بمباران اطلاعات بی‌فایده و نامرتبط همراه است. شناسایی اخبار مرتبط از میان حجم زیادی از مقالات خبری کار بسیار دشواری است. یک موتور جستجوی هوشمند اخبار، نه تنها باید تمام اسناد مرتبط را در مورد یک رویداد خاص بازیابی کند، بلکه یک دیدگاه کلی در مورد این که این رویداد چگونه ریشه می‌گیرد و تکامل می‌یابد، ارائه می‌دهد [۱]. بنابراین سایت‌های گردآورنده اخبار مانند Yahoo، Google و AOL سایت‌های گردآوری اخبار را برپا کرده‌اند که سرخط اخبار را از هزاران منبع خبری چندزبانه در سراسر جهان جمع‌آوری می‌کنند. هدف این سایت‌ها کمک به خوانندگان برای یافتن خبرهای جدید و جالب است. برخی از آنها ویژگی‌هایی فراتر از تجمیع اخبار، برای افزایش ارزش خدمات خود ارائه می‌دهند. یکی از این ویژگی‌ها، شخصی‌سازی محتوای نشان داده شده به هر کاربر است [۲].

یک موضوع[1]، مجموعه‌ای از اسناد مرتبط به یک رویداد است که زمان و مکان مشخصی دارد. برای نمونه، افتتاح یک استادیوم بزرگ در تهران می‌تواند باعث پیدایش گزارش‌های خبری تکراری در رسانه‌های خبری ایران و جهان شود. حتی یک موضوع بر اساس اهمیت، ممکن است در طول چند روز یا حتی چند هفته در یک رسانه منتشر شود. بنابراین یک خوشه از مقالات خبری در مورد یک موضوع، مزیت‌های عملی بسیاری برای خوانندگان اخبار دارد. از آنجا که هیچ کاربری نمی‌تواند تمام اخبار منتشرشده در جهان را بخواند، اغلب مردم به طور طبیعی به اخبار مربوط به موضوعات مورد علاقه‌شان توجه می‌کنند. به عنوان مثال، مدیران اقتصادی ممکن است به هر موضوعی مربوط به نوسانات نرخ طلا و ارز علاقه‌مند باشند، در حالی که نوجوانان تنها به جدیدترین مدها گرایش دارند [۳].

۲- بیان مسأله

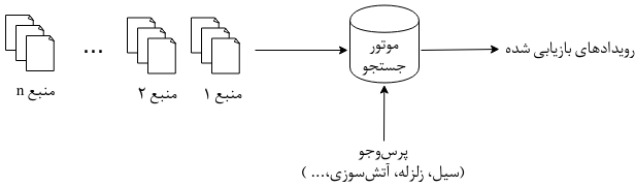
هدف از این پژوهش ارائه یک موتور جستجو برای تشخیص و خلاصه‌سازی رویدادها از جریان اسناد خبری می‌باشد. تشخیص رویداد، وظیفه کشف و گروه‌بندی اسنادی را دارد که رویدادی یکسان را شرح می‌دهند و با ارائه یک ساختار مفهومی از گزارش‌های خبری، هدایت بهتر کاربران در فضاهای خبری را تسهیل می‌کند [۴]. به طور دقیق‌تر، با داشتن جریان اسناد متنی منتشرشده در یک دوره زمانی مشخص، هدف اصلی تشخیص رویداد، تحلیل اسناد و استخراج مجموعه رویدادهایی است که در طول دوره در جهان رخ داده است. یک رویداد به صورت غیر دقیق به عنوان چیزی که در یک زمان خاص و در یک مکان خاص اتفاق می‌افتد، تعریف می‌شود [۵]. از دیدگاه داده‌کاوی، تشخیص رویداد ممکن

این مقاله در تاریخ ۲۱ آذر ماه ۱۳۹۹ دریافت و در تاریخ ۲۷ تیر ماه ۱۴۰۰ بازنگری شد.

علیرضا میرزائیان، دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران، (email: a.mirzaeiyan@mail.sbu.ac.ir).

صادق علی‌اکبری (نویسنده مسئول)، دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران، (email: s_aliakbary@sbu.ac.ir).

1. Topic



شکل ۱: نمای کلی سامانه پیشنهادی.

است شبیه خوشه‌بندی سند یا کلمه باشد. اگرچه تعدادی از مشخصه‌های اصلی تشخیص رویداد وجود دارند که اگر در نظر گرفته نشوند، می‌توانند بر عملکرد خوشه‌بندی تأثیر منفی بگذارند:

- زمان، نقشی محوری در هر سند خبری ایفا می‌کند.
- موضوعات خبری به طور طبیعی انفجاری هستند.
- اسناد خبری با محتوای مشابه معنایی اما جدا از فاصله زمانی با فریم‌های زمانی متفاوت به احتمال زیاد از موضوعات مختلف نشأت گرفته‌اند.

ما در این تحقیق روشی گذشته‌نگر را پیشنهاد می‌کنیم که مبتنی بر بازنمایی رویدادها به شکل کلمات کلیدی است. سامانه پیشنهادی قادر است با جمع‌آوری و نمایه‌سازی[1] اسناد خبری از منابع گوناگون و دریافت یک پرس‌وجو از کاربر، رویدادها را بازیابی کند. شکل ۱ نمایی کلی از سامانه پیشنهادی را نمایش می‌دهد.

روش پیشنهادی با دریافت یک بازه زمانی و یک پرس‌وجو مثل زلزله، سیل و یا انفجار که به شکلی رویداد را توصیف می‌کند، تمام رویدادهای مرتبط به پرس‌وجو را که در آن بازه زمانی اتفاق افتاده‌اند بازیابی می‌کند. در سامانه پیشنهادی، هر رویداد به شکل یک بازه زمانی و تعدادی کلمه کلیدی مرتبط بازنمایی می‌شود تا یک خلاصه سطح بالا از رویداد ارائه شود. فرض روش پیشنهادی این است که کلمات مرتبط که اغلب در طول یک دوره زمانی با هم تکرار می‌شوند، نماینده رویدادهایی مشابه هستند که در آن زمان اتفاق افتاده است [۵]. نهایتاً با استفاده از کلمات و بازه زمانی استخراج‌شده، اسناد مرتبط به رویداد، بازیابی و خلاصه‌سازی می‌شوند و بدین ترتیب کاربر می‌تواند به راحتی یک تاریخچه کلی از رویدادها را استخراج کند. چنین سامانه‌ای نه تنها برای کاربران معمولی بلکه برای روزنامه‌نگاران، تاریخ‌دانان و حتی دانشجویان مفید خواهد بود.

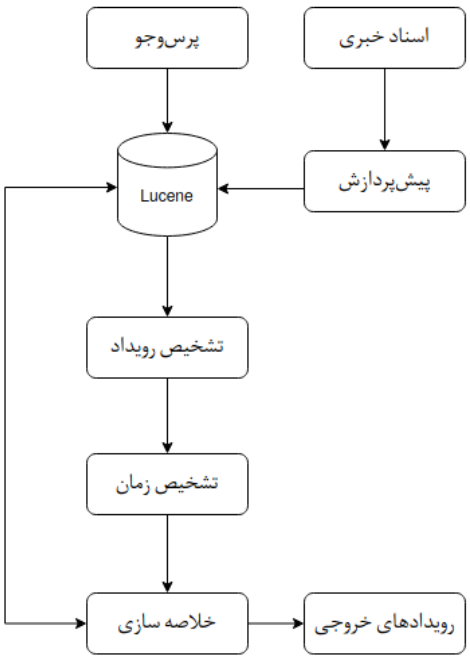
۳- کارهای پیشین

حوزه تحقیقاتی تشخیص رویداد، نشأت‌گرفته از تشخیص موضوع است که در سال ۱۹۹۷ تحت عنوان یک برنامه به نام تشخیص و ردیابی موضوع در آژانس پروژه‌های تحقیقاتی پیشرفته دفاعی آمریکا آغاز شد. انگیزه طرح تحقیقاتی تشخیص و ردیابی موضوع، ارائه فناوری اصلی برای ابزارهای نظارت بر اخبار موجود در منابع متعدد رسانه‌های سنتی برای مطلع‌ساختن کاربران در مورد اخبار و تحولات بود [۶]. یکی از روش‌های قدیمی در تشخیص موضوع، مدل‌سازی موضوعی می‌باشد که از مدل‌های آماری برای شناسایی رویدادها به عنوان متغیرهای پنهان در اسناد استفاده می‌کند. تخصیص دیریکله نهفته (LDA)[2] یک مدل برای تحلیل متن است که به طور گسترده مورد استفاده قرار می‌گیرد [۷]. هی و همکاران [۸] مسئله تجزیه و تحلیل خط سیر در دامنه‌های زمان و فرکانس را با هدف شناسایی کلمات مهم و کمتر گزارش‌شده و کلمات متناوب و نامتناوب مورد بررسی قرار دادند. در این روش فرکانس سند هر کلمه مانند یک سری زمانی در نظر گرفته می‌شود که فرکانس سند را در نقاط زمانی مختلف نشان می‌دهد. سپس برای دسته‌بندی ویژگی‌ها به کلمات مهم و کمتر گزارش‌شده و کلمات متناوب و نامتناوب از آنالیز طیفی استفاده می‌شود. در این روش ویژگی‌های نامتناوب را با چگالی گاوسی و ویژگی‌های متناوب را با چگالی‌های ترکیبی گاوسی مدل‌سازی کردند و متعاقباً انفجار هر ویژگی را با پارامترهای این مدل‌ها شناسایی نمودند و از یک الگوریتم تشخیص رویداد حریصانه و بدون نظارت برای شناسایی هر دو رویداد متناوب و نامتناوب استفاده کردند. در کار مشابه دیگر، کالا [۵] با الهام از هی و همکاران یک سامانه گذشته‌نگر پیشنهاد کرد که با استفاده از ویژگی‌های زمانی و معنایی، کلمات را خوشه‌بندی می‌کند. در این روش ابتدا یک ماتریس فاصله ایجاد می‌شود که درایه‌های آن فاصله بین کلمات را از لحاظ ویژگی‌های زمانی و فاصله زمانی نشان می‌دهد، سپس با استفاده از روش خوشه‌بندی مبتنی بر چگالی این کلمات خوشه‌بندی می‌شوند. سومیکاوا و همکارانش [۹] یک سامانه جستجوی تعاملی برای جستجوی رویدادهای تاریخی با فیلترینگ مبتنی بر دسته‌بندی پیشنهاد کردند که از طریق رده‌بندی خودکار رویداد محقق شده است.

این سامانه چهار نوع داده ورودی دارد: پرس‌وجوی متنی، دسته‌بندی رویداد، محدوده زمانی و روش رتبه‌بندی. خروجی این سامانه فقط شامل شرح رویدادهایی است که دارای کلمات پرس‌وجو هستند و همچنین در بازه زمانی ورودی منتشرشده باشند. در کار مشابه دیگر متزلر و همکارانش [۱۰] مسئله بازیابی ساختارمند اطلاعات رویدادهای تاریخی از بایگانی‌های میکروبلاگ را با پیشنهاد یک تکنیک گسترش پرس‌وجوی زمانی جدید حل کردند. هو و همکاران [۴] روشی را پیشنهاد کردند که در آن اسناد بر پایه تعبیه کلمات بازنمایی می‌شوند. در روش پیشنهادی هو، به طور خاص ابتدا یادگیری تعبیه کلمات انجام می‌شود و سپس با استفاده از الگوریتم خوشه‌بندی، کلمات در کلاس‌های معنایی مختلف دسته‌بندی می‌شوند. هر سند خبری به صورت یک توزیع روی کلاس‌های معنایی بازنمایی می‌شود و نهایتاً با استفاده از یک الگوریتم خوشه‌بندی تطبیقی برخط رویدادها تشخیص داده می‌شوند. موتیدیس و همکاران [۱۱] یک روش مبتنی بر شبکه پیشنهاد کردند که فرض آن بر این است که رویدادهای خبری مهم همیشه شامل موجودیت‌های نامدار (مانند افراد، اماکن و سازمان‌ها) است که در مقالات خبری با هم مرتبط هستند. در این روش یک سری شبکه‌های زمانی ایجاد می‌شود که در آنها موجودیت‌های شناسایی‌شده بر اساس همنشینی در مقالات و جملات به هم متصل هستند. در این روش، درجه گره وزن‌دار در طول زمان ردیابی شده و تشخیص نقطه تغییر، جهت تعیین موقعیت وقایع مهم مورد استفاده قرار می‌گیرد. رویدادهای بالقوه با استفاده از تشخیص انجمن‌ها در گراف کلمات که موجودیت‌های نام‌دار و عبارات اسمی را به هم متصل می‌کند مشخص و متمایز می‌شوند.

با مرور کارهای پیشین در این حوزه، کمبود نسبی منابع عمومی برای ارزیابی عملکرد روش‌های تشخیص، بازیابی و خلاصه‌سازی رویداد مشخص می‌شود. بنابراین ما در این تحقیق به دنبال ارائه یک سامانه مدولار هستیم که در آن می‌توان روش‌های مختلف مثل ویژگی محور، سند محور و مدل‌سازی موضوعی را ارزیابی کرد. علاوه بر این یک مسئله رایج در روش‌های تشخیص رویداد مشخص‌نبودن تعداد رویدادها است. بنابراین یکی از اهداف اصلی این تحقیق، ارائه یک روش جدید خوشه‌بندی است که نیازی به پارامتر پیشین تعداد خوشه ندارد و نسبت به

1. Indexing
2. Latent Dirichlet Allocation



شکل ۲: معماری سامانه بازیابی رویداد.

روش‌های دیگر کارایی بالاتری دارد. برخلاف عمده روش‌های گذشته‌نگر که کل اسناد را تحلیل می‌کنند، روش پیشنهادی، پرس‌وجو محور است و با محدودکردن موضوعات به پرس‌وجوی کاربر، تاریخچه رویدادهای مرتبط را تشکیل می‌دهد.

٤- روش پیشنهادی

سامانه پیشنهادی نیز یک سامانه گذشته‌نگر و غیر نظارتی[1] است که با دریافت یک پرس‌وجو، رویدادهای تاریخی را که در دنیای واقعی رخ داده‌اند از مجموعه داده خبری بازیابی می‌کند. شکل ۲ معماری سامانه را نمایش می‌دهد. در این نمودار چند پیمانه مهم وجود دارد که در ادامه به صورت مفصل به آنها پرداخته شده است.

در مرحله پیش‌پردازش، اسنادی که از منابع خبری مختلف جمع‌آوری شده‌اند برای پردازش در مراحل بعد آماده می‌شوند. در ابتدا اسنادی که تاریخ انتشار، عنوان و یا متن ندارند شناسایی گردیده و از پردازش‌های مراحل بعد کنار گذاشته می‌شوند. واحدسازی[2]، حذف ایست‌واژه‌ها و ریشه‌یابی[3] مراحلی هستند که برای نرمال‌سازی متن انجام می‌شود. پس از فرایند نرمال‌سازی، این متون با استفاده از ابزار متن‌باز لوسین[4] نمایه‌سازی می‌شوند.

کلیدی‌ترین پیمانه در سامانه پیشنهادی، تشخیص رویداد است. با توجه به پرس‌وجو محور بودن سامانه پیشنهادی و حجم زیاد اسناد موجود، برای محدودکردن ورودی این پیمانه، اسنادی که بیشترین ارتباط را به پرس‌وجوی کاربر دارند از نمایه بازیابی می‌شوند. سامانه پیشنهادی، مبتنی بر بازنمایی رویدادها به صورت مجموعه‌ای از کلمات کلیدی است. هر رویداد با یک مجموعه از کلمات کلیدی مرتبط تعریف می‌شود. در این پژوهش با توجه به روش‌هایی که از حوزه بازیابی اطلاعات و تحلیل شبکه استخراج گردیده است، یک روش بدون نظارت برای شناسایی رویدادهای خبری در مقالات خبری ارائه شده است. ما یک رویکرد مبتنی بر خوشه‌بندی را توسعه می‌دهیم که فرض آن بر این است که رویدادهای خبری مهم، همیشه شامل کلمات کلیدی است که در مقالات خبری با هم در یک بازه زمانی مشخص اتفاق می‌افتند. یک ویژگی مهم این روش این است که نیاز به هیچ فرض قبلی در مورد تعداد رویدادها در مجموعه اسناد خبری ندارد که خود یک بهبود کلیدی در روش‌های بدون نظارت موجود است. روش پیشنهادی از چهار مرحله اصلی تشکیل گردیده که در ادامه به آنها پرداخته شده است.

٤-١ تشخیص کلمات کلیدی

در ابتدا، مهم‌ترین کلمات در هر مقاله در مقایسه با کل پیکره زبانی شناسایی و رتبه‌بندی می‌شوند. به همین منظور، همه کلمات بر اساس فرمول TFIDF [۱۲] به صورت نزولی مرتب می‌گردند و مهم‌ترین کلمات برای پردازش در مراحل بعد استفاده می‌شوند. این روش امتیازدهی شدیداً کلمات غیر معمول در هر سند را انتخاب می‌کند. اگرچه این روش در بازیابی اطلاعات نسبتاً ساده است، اما از لحاظ نظری برای خوشه‌بندی رویداد بسیار مناسب است. شهود استفاده از این روش امتیازدهی این است که با پیداکردن غیر معمول‌ترین کلمات در هر مقاله در مقایسه با کل منابع خبری، این امکان وجود دارد که برخی از کلمات برجسته و مرتبط به رویداد مثل اشخاص، مکان‌ها و سایر کلمات مرتبط به رویداد را در هر مقاله استخراج کنیم. در مرحله بعد ویژگی‌های زمانی این کلمات بررسی می‌شوند تا کلمات رویدادخیز شناسایی شوند. به همین منظور فرکانس رخداد کلمات در طول زمان بررسی می‌گردد و بنابراین به ازای هر کلمه، یک خط سیر[5] تشکیل می‌شود. در تعریف هی و همکاران [۸] بازنمایی برداری خط سیر کلمه f به صورت (۱) تعریف می‌شود

$$y_f = [y_f(1), y_f(2), \ldots, y_f(T)] \qquad (1)$$

در این بردار $y_f(t)$ مقدار کلمه f در زمان t را نشان می‌دهد که این مقدار با امتیاز DFIDF تعریف می‌شود

$$y_f(t) = \frac{DF_f(t)}{N(t)} \times \log \frac{N}{DF_f} \qquad (2)$$

در این فرمول، $DF_f(t)$ تعداد اسناد روز t است که ویژگی f را دارند، DF_f تعداد اسنادی است که ویژگی f را در جریان T دارند، $N(t)$ تعداد کل اسناد روز t است و N تعداد کل اسناد را در جریان T نشان می‌دهد. DFIDF تغییریافته فرمول معروف TFIDF است که میزان اهمیت کلمه در مجموعه اسناد را نشان می‌دهد. هدف این تغییر اضافه‌کردن اطلاعات زمان و اندازه‌گیری اهمیت کلمه در طول زمان است [۵].

بعد از محاسبه خط سیر، به منظور تشخیص رویدادخیز بودن کلمه، میانگین پنجره‌های کشویی[6] به طول X محاسبه می‌شود. سپس کلماتی که مقدار DFIDF آنها از آستانه‌ای به اندازه Y واحد فاصله از میانگین حرکتی[7] بیشتر شود به عنوان کلمات رویدادخیز انتخاب می‌شوند. در آزمایش‌های انجام‌شده پنجره زمانی به طول یک هفته $(X = ۷)$ و آستانه به فاصله ۲ واحد از میانگین حرکتی $(Y = ۲)$ تعریف شده است.

1. Unsupervised
2. Tokenization
3. Stemming
4. Lucene
5. Trajectory
6. Sliding Window
7. Moving Average

Algorithm 2: Event Merging.

Input: Event sets $\{\mathbf{E}_t\}$, where $\mathbf{E}_t = \{E_{ti}\}$ contains a set of events in the t-th time-slice, Merging Threshold Δ
Output: Event set $\mathbf{E}=\{E_1, E_2, \ldots, E_K\}$

let $\mathbf{E} = \mathbf{E}_1$
foreach *event set* $\boldsymbol{E}_t$ *(t=2,… ,T)* **do**
  foreach *event* $E_{ti} \in \boldsymbol{E}_t$ **do**
    foreach *event* $E_k \in \boldsymbol{E}$ **do**
      let c_{ti}, c_k represent the centroid of E_{ti} and E_k respectively
      calculate the similarity $sim(c_{ti}, c_k)$
      let $maxS = \max_i sim(c_{ti}, c_k)$
      let $maxE = E_k | sim(c_{ti}, E_k) = maxS$
      if $maxS \geq \Delta$ **then**
        let $maxE \leftarrow E_{tj} \cup maxE$
        recalculate the representation of $maxE$ by the centroid
      else
        let $E_{ti} \in \mathbf{E}$
return all event-centric clusters $\mathbf{E} = \{E_1, E_2, \ldots E_K\}$

شکل ۳: الگوریتم ادغام رویدادها [۴].

۴-۲ تشکیل شبکه

بعد از استخراج کلمات رویدادخیز از مجموعه اسناد، برای n کلمه کلیدی تشخیص داده شده یک ماتریس شباهت $n \times n$ تشکیل می‌دهیم که درایه‌های آن میزان همبستگی دوبه‌دو بین تمام کلمات را بازنمایی می‌کند. برای اندازه‌گیری همبستگی معنایی بین دو کلمه w_i و w_j از شباهت کسینوسی بردار تعبیه لغات استفاده می‌شود

$$\cos(w_i, w_j) = \frac{w_i . w_j}{\|w_i\| . \|w_j\|} \qquad (۳)$$

مرتبه اجرایی تولید این ماتریس شباهت $O(n^2)$ است و برای ابعاد بزرگ، زمان اجرای تولید این ماتریس ممکن است چند روز یا چند هفته طول بکشد که این مسئله می‌تواند یک مانع مهم برای کار تحقیق باشد. یک راه ساده برای کاهش این پیچیدگی، تولید این ماتریس برای پنجره‌های زمانی متحرک است [۱۳]. از طرفی در تشخیص رویداد، اطلاعات زمانی بسیار مهم است. پس از وقوع رویداد، تعداد زیادی از اسناد خبری در زمان کوتاهی منتشر می‌شوند. اسناد خبری با فواصل زمانی طولانی‌تر معمولاً در مورد وقایع مختلف هستند. بنابراین تقسیم اسناد خبری با توجه به اطلاعات زمان انتشار، برای شناسایی اسناد خبری که رویدادی یکسان را گزارش می‌کنند مفید است [۴]. در روش پیشنهادی، ما اسناد خبری را به ترتیب زمانی مرتب می‌کنیم و آنها را به خوشه‌هایی با اندازه پنجره زمانی ثابت تقسیم می‌نماییم. در این پژوهش طول هر پنجره زمانی سه روز در نظر گرفته شده است. پس از تقسیم جریان اسناد خبری، به خوشه‌هایی می‌رسیم که هر کدام با یک برش زمانی متناظر است. بنابراین ما از کلمات رویدادخیز موجود در هر برش زمانی برای تشکیل یک ماتریس شباهت مجزا استفاده می‌کنیم تا هم زمان اجرا کاهش یابد و هم فرض رخداد کلمات کلیدی مرتبط در یک بازه زمانی لحاظ شود. در نهایت ماتریس شباهت ساخته‌شده را می‌توان به یک شبکه شباهت تبدیل کرد چرا که شبکه‌ها ذاتاً در مقایسه‌های دوبه‌دو بین اشیا به وجود می‌آیند. حاصل این تبدیل، شبکه‌ای وزن‌دار است که گره‌های آن را کلمات کلیدی و یال‌های آن را مقادیر شباهت بین کلمات تشکیل می‌دهند.

۴-۳ تشخیص انجمن

در این مرحله با استفاده از تکنیک‌های تشخیص انجمن[1]، شبکه تشکیل‌شده را خوشه‌بندی می‌کنیم. خروجی این خوشه‌بندی مجموعه‌ای از خوشه‌های منسجم از کلمات است که هر کدام یک رویداد را بازنمایی می‌کند. مسئله دریافت بازنمایی مبتنی بر شبکه از داده‌ها و تخصیص نودهای شبکه به گروه‌ها، تشخیص انجمن نامیده می‌شود [۱۳]. الگوریتم‌های زیادی برای تشخیص انجمن وجود دارد که در این بین ما از روش لوون[2] استفاده می‌کنیم. این روش امتیاز پیمانگی[3] را برای هر انجمن بیشینه می‌کند. پیمانگی، کیفیت گره‌های تخصیص‌یافته به انجمن‌ها را تعیین می‌کند. در واقع این روش ارزیابی می‌کند که گره‌های یک انجمن در مقایسه با یک شبکه تصادفی چه قدر می‌تواند منسجم‌تر باشد. این روش برای شبکه‌های غیر جهت‌دار و وزن‌دار مورد استفاده قرار می‌گیرد [۱۴]. البته، روش پیشنهادی مدولار است و استفاده از دیگر پیاده‌سازی‌ها ممکن است. برای مثال، این امکان وجود دارد که رویکرد محاسبه تشابه دوبه‌دو برای تشکیل شبکه با هر معیار مشابه دیگری جایگزین شود. به همین ترتیب، تعداد زیادی از روش‌ها در علوم کامپیوتر و شبکه‌های اجتماعی برای خوشه‌بندی شبکه‌های پیچیده وجود دارد که می‌تواند برای شبکه شباهت اعمال شود.

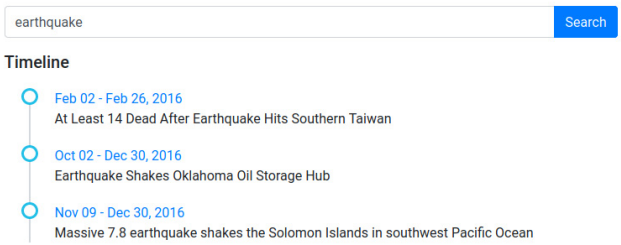
۴-۴ ادغام رویدادها

بعد از تشخیص رویدادها در هر برش زمانی، ما رویدادهایی را که به یک رویداد اشاره دارند اما در برش‌های متفاوتی هستند، ادغام می‌کنیم. با این روش می‌توان دقت خوشه‌های به دست آمده را بهبود بخشید. هو و همکاران [۴] یک روش سندمحور برای تشخیص رویدادهای خبری پیشنهاد کردند که پس از خوشه‌بندی، خوشه‌های به دست آمده را بر اساس شباهت معنایی ادغام می‌کند. ما نیز از همین روش برای ادغام رویدادهای به دست آمده استفاده می‌کنیم با این تفاوت که خروجی سامانه پیشنهادی ما ویژگی‌محور است. شکل ۳ الگوریتم پیشنهادی هو و همکاران را برای ادغام رویدادها در برش‌های زمانی متفاوت نشان می‌دهد. در این الگوریتم به صورت افزایشی رویدادهای هر برش زمانی $E(t) = \{Eti\} (t = 2, \ldots, T)$ با رویدادهای قبلی E (در ابتدا، $E = E_1$) ادغام می‌شوند. هر رویداد E با بردار مرکزوار خوشه متناظر ($c = \sum wi / N$) بازنمایی می‌شود. هر رویداد با مشابه‌ترین رویداد قبلی که شباهت بین آنها از یک آستانه از پیش تعریف شده Δ بیشتر باشد ادغام می‌شود. خروجی این الگوریتم تعدادی رویداد $E = \{E_1, E_2, \ldots, E_k\}$ است که به رویدادی یکسان اشاره نمی‌کنند.

۴-۵ تشخیص زمان

برای تشخیص زمان رویداد ابتدا باید خط سیر آن را از روی خط سیر کلمات کلیدی بسازیم. این کار را با محاسبه میانگین خطوط سیر کلمات کلیدی رویداد انجام می‌دهیم. سپس با استفاده از میانگین حرکتی و آستانه تعریف‌شده در پیمانه تشخیص رویداد، نویزهای خط سیر را حذف می‌کنیم. هی و همکاران [۸] برای تشخیص بازه زمانی رویدادهای متناوب و نامتناوب مدل‌های احتمالاتی متفاوتی را ارائه کرده‌اند. برای استفاده از این مدل‌ها ابتدا باید نوع رویداد را مشخص کنیم. هدف سامانه

1. Community Detection
2. Louvain
3. Modularity



شکل ۴: رابط کاربری سامانه پیشنهادی.

پیشنهادی، تشخیص رویدادهای تاریخی نامتناوب است که در یک زمان مشخص رخ داده‌اند. رویدادهای متناوب شامل کلماتی هستند که در بازه‌های زمانی مشخص تکرار می‌شوند، برای مثال کلمات مرتبط به پیش‌بینی آب و هوا که هر روز اعلام می‌شود یا کلمات مرتبط به مسابقاتی که اواخر هر هفته انجام می‌شود. بنابراین ما ابتدا رویدادهای نامتناوب را تشخیص می‌دهیم و سپس از مدل احتمالاتی ارائه‌شده برای تشخیص بازه زمانی رویداد استفاده می‌کنیم.

هی و همکاران [۸] خط سیر کلمات را به عنوان سیگنال‌های زمانی تفسیر می‌کردند که به آنها اجازه می‌داد تا خط سیر را با استفاده از تکنیک‌های پردازش سیگنال تحلیل کنند و کلمات متناوب را از غیر متناوب تشخیص دهند. ما تبدیل فوریه گسسته را برای نشان‌دادن سری‌های زمانی به صورت ترکیب خطی از امواج مختلط روی خط سیر اعمال می‌نماییم و $Fy_w=[X_1,X_2,\dots,X_T]$ را با استفاده از رابطه زیر محاسبه می‌کنیم

$$X_k=\sum_{t=1}^{T} y_w(t)e^{-\frac{2\pi i}{T}(k-1)t},\ k=1,2,\dots,T \qquad (۴)$$

خط سیر اصلی را می‌توان با فرکانس‌های غالب بازسازی کرد که این فرکانس‌ها را می‌توان از طیف توان[1] با استفاده از برآوردگر معروف تناوب‌نگار[2] تعیین نمود. این برآوردگر دنباله‌ای از مقدار مربع ضرایب فوریه است، $\|X_k\|^2, k=1,2,\dots,[T/2]$ که نشان‌دهنده قدرت سیگنال در فرکانس k/t در طیف است. برای تشخیص تناوب خط سیر نیاز به مشخص‌کردن دوره غالب[3] می‌باشد. بعد از محاسبه تناوب‌نگار، دوره غالب به عنوان معکوس فرکانس مربوط به بالاترین نقطه در طیف توان تعریف می‌شود

$$DP=\frac{T}{\arg\max_{k\le[\frac{T}{2}]}\|X_k\|^2} \qquad (۵)$$

سپس رویدادهایی را که فقط یک بار در جریان رخ داده‌اند (یعنی $DP>[T/2]$) نامتناوب در نظر می‌گیریم. به طور مشابه رویدادهایی که بیشتر از یک بار رخ داده‌اند ($DP\le[T/2]$) متناوب هستند. مشابه مدل ارائه‌شده توسط هی و همکاران [۸]، ما برای تشخیص بازه زمانی رویدادهای نامتناوب، خط سیر رویدادها را با توزیع گاوسی[4] برازش می‌کنیم و دو پارامتر میانگین و انحراف معیار را تخمین می‌زنیم. نهایتاً با استفاده از تابع چگالی احتمال برازش‌شده، بازه رویداد را به عنوان مناطقی با بیشترین چگالی به صورت $[\mu-\sigma,\mu+\sigma]$ تعریف می‌کنیم.

1. Power Spectrum
2. Periodogram
3. Dominant Period
4. Gaussian Distribution

۴-۶ خلاصه‌سازی

پس از تشخیص رویداد، همچنان نیازمند به تعریف رویدادها در قالبی هستیم که برای کاربر قابل درک باشد. مجموعه کلمات کلیدی ممکن است یک نمایش دقیق برای کامپیوتر باشد، اما بینش زیادی در مورد خود رویداد ارائه نمی‌دهد. بنابراین برای این کلمات باید یک شرح متنی تولید شود تا کاربر با خواندن آن بتواند درک درستی از رویداد به دست آورد و در صورت نیاز اسناد بیشتری را مطالعه کند.

یک روش ساده برای خلاصه‌سازی یک رویداد، بازیابی سرخط مرتبط‌ترین سند خبری از لحاظ تشابه حرکتی کلمه[5] است [۵]. فاصله حرکتی کلمه با استفاده از مدل Word۲Vec، شباهت دو سند را به صورت حداقل فاصله بردار کلمات یک سند برای رسیدن به بردار کلمات سند دوم اندازه‌گیری می‌کند [۱۵] و این تابع فاصله ترتیب کلمات را لحاظ نمی‌کند، به همین دلیل برای پرسش‌های کلمات کلیدی ما مناسب است. این روش بهترین نتایج را برای اسناد کوتاه به دست می‌آورد چون از نظر محاسباتی برای متون طولانی‌تر پرهزینه است و به همین دلیل ما این روش را برای سرخط خبری اسناد اعمال می‌کنیم. در چارچوب [۱۶] که فاصله حرکتی کلمه را پیاده‌سازی می‌کند، فرمول شباهت حرکتی کلمه به صورت مقابل تعریف می‌شود

$$Sim_{wmd}(d_i,d_j)=\frac{1}{1+wmd(d_i,d_j)} \qquad (۶)$$

در این فرمول اگر $WMD(d_i,d_j)=0$ باشد، مقدار این تابع ۱ و اگر $WMD(d_i,d_j)\to\infty$ این مقدار به ۰ نزدیک می‌شود. بنابراین برای خلاصه‌سازی رویدادهای نامتناوب بازیابی‌شده، ابتدا اسناد مرتبط به پرس‌وجوی کاربر را که در بازه زمانی رویداد منتشر شده‌اند از نمایه بازیابی می‌کنیم و سپس آنها را بر اساس امتیاز (۶) به صورت نزولی مرتب می‌نماییم و اولین سرخط را به عنوان خلاصه رویداد گزارش می‌کنیم. شکل ۴ رویدادهای خروجی سامانه پیشنهادی پس از اجرای مرحله خلاصه‌سازی را نمایش می‌دهد.

۵- روش ارزیابی

برای ارزیابی روش پیشنهادی از مجموعه داده مقالات خبری All The News که در وب‌سایت Kaggle موجود است استفاده می‌کنیم. این مجموعه داده شامل تقریباً دویست هزار مقاله از ۱۵ منبع خبری اصلی در ایالات متحده آمریکا است. برای ارزیابی از سه معیار رایج در بازیابی اطلاعات یعنی دقت[6]، یادآوری[7] و F-Measure استفاده می‌شود. به همین جهت، یک مرجع درستی[8] شامل ۳۲ پرس‌وجو و نزدیک به ۱۷۸ رویداد طراحی شده است. اطلاعات مربوط به رویدادهای مرجع، شامل زمان رویداد و شرح متنی آن از وب‌سایت OnThisDay استخراج شده است. این وب‌سایت با بیش از دویست هزار رویداد ثبت‌شده یک مرجع معتبر برای رویدادهای تاریخی محسوب می‌شود و با ارائه یک سامانه جستجوی تعاملی، امکان بازیابی رویدادها از پایگاه داده‌ای که توسط نویسندگان مختلف توسعه پیدا کرده است را فراهم می‌کند. برای ایجاد مرجع درستی، هر پرس‌وجو که شامل یک کلمه کلیدی می‌شود به همراه بازه زمانی

5. Word Mover's Similarity
6. Precision
7. Recall
8. Ground Truth

جدول ۱: نتایج ارزیابی دقت، یادآوری و F-Measure.

روش‌ها	دقت	یادآوری	F-Measure
KMeans	۰/۲۷۶۸	۰/۱۶۲۸	۰/۲۰۵۱
LDA	۰/۳۱۸۰	۰/۲۱۷۵	۰/۲۵۸۳
FPM	۰/۳۳۵۷	۰/۲۴۵۴	۰/۲۸۳۵
DBSCAN	۰/۳۹۷۵	۰/۳۲۸۹	۰/۳۶۰۰
Our Method	۰/۴۱۷۳	۰/۳۷۴۸	۰/۳۹۴۹

جدول ۲: نتایج ارزیابی زمان اجرا.

روش‌ها	زمان اجرا (ثانیه)
KMeans	۲۳۲
LDA	۱۳۵۲
FPM	۳۳۲
DBSCAN	۲۶۰۰
Our Method	۴۳۸

مجموعه داده به سامانه OnThisDay داده شده و رویدادهای بازیابی‌شده از این سامانه به عنوان مرجع رویدادهای پرس‌وجو برای ارزیابی استفاده می‌شود.

اگرچه فهرست مرجع رویدادهای حقیقی کامل نیست، اما این مرجع یک راه برای مقایسه عینی با روش‌های تشخیص رویداد ارائه می‌دهد. در حالی که رویدادهای بیشتری وجود دارند که در طول دوره زمانی بررسی‌شده رخ می‌دهند، نمونه‌ای از وقایع به اندازه کافی مهم، یک تصویر از کارایی روش‌ها را فراهم می‌کند [۵]. در ارزیابی صورت‌گرفته، رویدادها از نظر شباهت معنایی و زمانی بررسی شده‌اند. در صورتی که در بازه زمانی رویداد تشخیص داده شده، رویدادی در مرجع درستی وجود داشته باشد، بیشینه شباهت کسینوسی میانگین بردار تعبیه لغات رویداد تشخیص داده شده و لغات رویدادهای متناظر در مرجع درستی به عنوان وزن رویداد در نظر گرفته می‌شود

$$Precision(Q) = \frac{1}{|Detected|} \times \sum_{i=1}^{Detected} \max\{\frac{D_i^H \times R^H}{\|D_i^H\| \times \|R^H\|} R \in Reference\} \quad (۷)$$

$$Recall(Q) = \frac{1}{|Reference|} \times \sum_{i=1}^{Reference} \max\{\frac{D^H \times R_i^H}{\|D^H\| \times \|R_i^H\|} D \in Detected\} \quad (۸)$$

در ارزیابی صورت‌گرفته، میانگین دقت و یادآوری پرس‌وجوها محاسبه و گزارش شده است. دو معیار ذکرشده به تنهایی نمی‌توانند کارایی سامانه را نشان دهند و معمولاً از میانگین هارمونیک دو معیار دقت و یادآوری استفاده می‌گردد که به آن F-Measure گفته می‌شود.

۶- تحلیل نتایج

ما عملکرد روش پیشنهادی را با چندین روش پایه بر اساس دو معیار دقت و زمان اجرا مقایسه کردیم. برای ادغام رویدادها، پارامتری را انتخاب کردیم که منجر به بهترین نتیجه در معیار دقت و یادآوری می‌شود. روش‌های پایه ارزیابی شامل موارد زیر است:

- KMeans: یک روش محبوب برای خوشه‌بندی در داده‌کاوی است. در پیاده‌سازی این روش هر سند با بردار TFIDF بازنمایی می‌شود. همچنین تعداد خوشه‌ها برابر با تعداد رویدادهای موجود به ازای هر پرس‌وجو تنظیم می‌شود.
- LDA: تخصیص دیریکله نهفته یک مدل موضوعی برای تحلیل متن است که به صورت گسترده استفاده می‌شود. تعداد موضوعات این روش برابر با تعداد رویدادهای هر پرس‌وجو تنظیم شده است.
- FPM: یک تکنیک معروف کاوش تراکنش برای شناسایی موضوع می‌باشد که مشخص می‌کند کدام کلمات در مجموعه تراکنش‌ها با هم اتفاق می‌افتند.
- DBSCAN: یک روش خوشه‌بندی ویژگی‌محور و مبتنی بر چگالی است که با به کارگیری ویژگی‌های زمانی و معنایی کلمات را خوشه‌بندی می‌کند [۵].

پارامترهای این روش‌ها به ازای مقادیر مختلف ارزیابی شده و بهترین نتایج در جدول ۱ گزارش شده است. نتایج جدول نشان می‌دهد که دقت و یادآوری در روش‌های مختلف نسبتاً پایین است. علت این موضوع همپوشانی نسبتاً پایین رویدادهای تشخیص داده شده و مرجع درستی می‌باشد. با این حال ارزیابی بر اساس نمونه‌ای از رویدادهای مهم می‌تواند به ما کمک کند تا تصویر درستی از کارایی روش‌ها به دست آوریم. نتایج نشان می‌دهد که روش پیشنهادی نسبت به روش‌های پایه در معیار دقت و یادآوری کارایی بالاتری دارد. همچنین روش‌های KMeans و LDA نسبت به سایر روش‌ها کارایی پایین‌تری دارند. علت این موضوع این است که این دو روش از اطلاعات زمان استفاده نمی‌کنند، در حالی که DBSCAN با بهره‌گیری از ویژگی‌های زمانی و معنایی کلمات را خوشه‌بندی می‌کند. روش‌های داده‌پردازی الگوهای تکراری و روش پیشنهادی نیز با استفاده از پنجره‌های زمانی متحرک جریان اسناد را تقسیم می‌کنند. در نتیجه استفاده از پنجره‌های زمانی متحرک می‌تواند برای تشخیص رویداد مفید باشد. تحقیقات گذشته نیز نشان می‌دهد که تقسیم اسناد خبری با توجه به اطلاعات زمان انتشار، برای شناسایی اسناد خبری که رویدادی یکسان را گزارش می‌کنند مفید است [۴].

در تشخیص رویداد، کارایی زمان اجرا بسیار مهم است. به همین منظور مدت زمان اجرای روش‌های پایه به همراه روش پیشنهادی ارزیابی شده است. برای ارزیابی زمان اجرا از یک کامپیوتر شخصی با حافظه ۱۶ GB و پردازشگر اینتل i۷-۷۵۰۰U استفاده شده است. جدول ۲ زمان اجرای روش‌های پایه و پیشنهادی را نشان می‌دهد. بررسی نتایج ارزیابی زمان اجرا در جدول ۲ نشان می‌دهد که روش KMeans و DBSCAN به ترتیب بهترین و بدترین زمان اجرا را دارند. روش پیشنهادی نیز از لحاظ زمان اجرا، هم‌رده با بهترین روش‌ها می‌باشد، هرچند در روش پیشنهادی پنجره‌ها به صورت ترتیبی پردازش می‌شوند و در صورت موازی‌سازی فرایند، می‌توان زمان اجرا را تا حدی کاهش داد.

۷- نتیجه‌گیری

در این پژوهش به بررسی تشخیص رویداد گذشته‌نگر از جریان‌های متنی پرداخته شد. در روش پیشنهادی، رویدادها با کلمات کلیدی توصیف می‌شوند که به لحاظ معنایی به هم مرتبط هستند و در یک بازه زمانی یکسان با هم اتفاق می‌افتند. این روش با به کارگیری یک روش خوشه‌بندی مبتنی بر شبکه، کلمات مرتبط به یک رویداد را که از لحاظ زمان و معنا با هم همبستگی دارند شناسایی می‌کند. در نهایت، زمانی که رویدادها شناسایی می‌شوند از کلمات کلیدی هر رویداد برای تولید یک توصیف متنی و یک بازه زمانی استفاده می‌شود. یکی از نوآوری‌های این

پژوهش، استفاده مؤثر از خوشه‌بندی داده‌ها جهت بازیابی دقیق رویدادها و تشخیص مناسب زمان رویدادها است.

در این پژوهش با ارائه یک سامانه مدولار، روش پیشنهادی و روش‌های پایه تشخیص رویداد، ارزیابی و مقایسه شدند. نتایج نشان می‌دهد که استفاده از روش خوشه‌بندی ارائه‌شده در معیار دقت و یادآوری نسبت به سایر روش‌ها کارایی بالاتری دارد. همچنین مقایسه روش‌ها نشان می‌دهد که استفاده از اطلاعات زمان مثل پنجره‌های زمانی متحرک می‌تواند برای تشخیص رویداد مفید باشد.

به عنوان کارهای تحقیقاتی آینده که می‌توانند جهت توسعه سامانه پیشنهادی در این تحقیق انجام شوند، پیشنهادهای زیر مطرح می‌شود:

- افزایش سرعت اجرا با استفاده از موازی‌سازی فرایند تشخیص رویداد
- به کارگیری مستقیم موجودیت‌های نام‌دار مثل اشخاص، مکان‌ها و سازمان‌ها جهت افزایش کارایی دقت تشخیص رویداد
- آزمایش روش‌های دیگر تشخیص انجمن در علم شبکه برای بهبود دقت روش خوشه‌بندی
- بهبود مدول خلاصه‌سازی با بهره‌گیری از روش‌های خلاصه‌سازی اسناد در پردازش زبان طبیعی
- ارائه معیارهای جدید همبستگی برای تشکیل شبکه کلمات

مراجع

[1] S. Lv, *et al*., "Yet another approach to understanding news event evolution," *World Wide Web*, vol. 23, no. 4, pp. 2449-2470, May 2020.

[2] O. N. N. Fernando and C. W. Chang, "Twittener: an aggregated news platform," in *Proc. IEEE Int. Conf. on Cyberworlds,* pp. 378-381, Kyoto, Japan, 2-4 Oct. 2019.

[3] Q. He, *Topical Analysis of Text Streams*, Ph.D. Dissertation, Nanyang Technological University, Singapore, 2009.

[4] L. Hu, B. Zhang, L. Hou, and J. Li, "Adaptive online event detection in news streams," *Knowledge-Based Systems,* vol. 138, pp. 105-112, 15 Dec. 2017.

[5] T. Kala, *Event Detection from Text Data*, Bacholor Thesis,Department of Cybernetics Faculty of Electrical Engineering, Czech Technical University in Prague, May 2017.

[6] F. Atefeh and W. Khreich, "A survey of techniques for event detection in twitter," *Computational Intelligence,* vol. 31, no. 1, pp. 132-164, Feb. 2015.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. of Machine Learning Research*, vol. 3, pp. 993-1022, Mar. 2003.

[8] Q. He, K. Chang, and E. P. Lim, "Analyzing feature trajectories for event detection," in *Proc. of the 30th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 207-214, Amsterdam, The Netherlands, 22-27 Jul. 2007.

[9] Y. Sumikawa and A. Jatowt, "System for category-driven retrieval of historical events," in *Proc. of the 18th ACM/IEEE on Joint Conf. on Digital Libraries*, pp. 413-414, Fort Worth Texas USA, 3-7 Jun. 2018.

[10] D. Metzler, C. Cai, and E. Hovy, "Structured event retrieval over microblog archives," in *Proc. of the Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 646-655, Montreal, Canada, 3-8 Jun. 2012.

[11] I. Moutidis and H. T. P. Williams, "Utilizing complex networks for event detection in heterogeneous high-volume news streams," *Complex Networks and Their Applications VIII: Proc. of the 8th Int. Conf. on Complex Networks and Their Applications*, vol. 1, pp. 659-672, Lisbon, Portugal, 10-12 Dec. 2019.

[12] H. Schutze, C. D. Manning, and P. Raghavan, *Introduction to Information Retrieval*, vol. 39, Cambridge University Press Cambridge, 2008.

[13] T. Nicholls and J. Bright, "Understanding news story chains using information retrieval and network clustering techniques," *Communication Methods and Measures, Routledge*, vol. 13, no. 1, pp. 43-59, 2019.

[14] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. of Statistical Mechanics: Theory and Experiment,* vol. 2008, no. 10, Article No.: P10008, Oct. 2008.

[15] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, "From word embeddings to document distances," in *Proc. of the 32nd Int. Conf. on Machine Learning*, vol. 37, pp. 957-966, Lille, France, 6-11 Jul. 2015.

[16] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proc. of LREC Workshop New Challenges for NLP Frameworks,* pp. 46-50, Valletta, Malta, 22-22 May 2010.

علیرضا میرزاییان مقطع کارشناسی را در دانشگاه زنجان در سال ۱۳۹۵ و مقطع کارشناسی ارشد را در دانشگاه شهیدبهشتی در سال ۱۳۹۹ به پایان برد. زمینه تحقیقاتی مورد علاقه ایشان عبارتند از: موتورهای جستجو و تشخیص رویداد از منابع خبری.

صادق علی‌اکبری مقطع کارشناسی مهندسی کامپیوتر را در دانشگاه تهران در سال ۱۳۸۵ به پایان برد. سپس در سال‌های ۱۳۸۷ و ۱۳۹۳ مقاطع کارشناسی ارشد و دکترا در همین رشته را در دانشگاه صنعتی شریف گذراند. وی از سال ۱۳۹۴ عضو هیأت علمی دانشکده مهندسی و علوم کامپیوتر دانشگاه شهید بهشتی بوده است. زمینه‌های علمی مورد علاقه نام‌برده عبارتند از: مهندسی نرم‌افزار مبتنی بر داده‌کاوی و تحلیل شبکه‌های پیچیده.