دانـــشگاه یـــزد

پردیس فنی و مهندسی

گروه مهندسی کامپیوتر

پایان‌نامه برای دریافت درجه کارشناسی ارشد

نرم افزار

تشخیص رویداد از وب‌سایت‌های خبری با استفاده از خوشه‌بندی

استادان راهنما: دکتر سجاد ظریف زاده و دکتر امیر جهانگرد رفسنجانی

پژوهش و نگارش: الهام رسولی

مهرماه ۱۳۹۷

کلیه‌ی حقوق مادی و معنوی مترتب بر نتایج مطالعات، ابتکارات و نوآوری‌های ناشی از تحقیق موضوع این پایان‌نامه/رساله متعلق به دانشگاه یزد است و هرگونه استفاده از نتایج علمی و عملی از این پایان‌نامه/رساله برای تولید دانش فنی، ثبت اختراع، ثبت اثر بدیع هنری، همچنین چاپ و تکثیر، نسخه‌برداری، ترجمه و اقتباس و ارائه مقاله در سمینارها و مجلات علمی از این پایان‌نامه/رساله منوط به موافقت کتبی دانشگاه یزد است.

تقدیم به

پدر و مادر عزیزم

و همه کسانی که درست اندیشیدن را به من آموختند.

سپاس‌گزاری

سپاس خداوند یکتای عزتمندی که رحمت و دانش او در سراسر گیتی گسترده شده، آسمان‌ها و زمین همه از آن اوست و علم و دانش حقیقی را بر هر که بخواهد موهبت می‌فرماید. رحمت و لطف او را بی‌نهایت سپاس می‌گویم چرا که فهم و درک مطالب این پژوهش را بر من ارزانی داشت و مرا به این اصل رساند که علم و ایمان دو بال یک پروازند. توفیق تلاش به من داد و هر بار که خطا کردم فرصتی دوباره، تا با امید، تلاشی تازه را آغاز کنم. سپاس بیکران بر همدلی و همراهی و همگامی پدر و مادر دلسوز و مهربانم و با تقدیر و تشکر شایسته از استادان فرهیخته جناب آقای دکتر سجاد ظریف زاده و جناب آقای دکتر امیر جهانگرد رفسنجانی که راهنمایی این پایان‌نامه را به عهده گرفتند و در تمام مراحل مرا از راهنمایی‌های ارزشمند خود بهره‌مند ساختند. همچنین از همه‌ی استادان دوران تحصیلم به ویژه جناب آقای دکتر علی‌محمد زارع بیدکی به خاطر تدریس فوق العاده‌ی ایشان و از در پایان همه‌ی دوستانم که در طی این مسیر با من همدل و همراه بودند کمال تشکر و قدردانی را دارم.

چکیده

با افزایش تولید محتوا و اسناد الکترونیکی در وب نیاز به ابزارهای استخراج و پالایش اطلاعات بیشتر احساس می‌شود. چرا که یافتن اطلاعات مورد نظر از میان انبوه اطلاعات کاری دشوار و زمان‌بر است. از این رو، ابزارهای مختلفی جهت تسهیل دسترسی به اطلاعات مورد نظر کاربران ایجاد شده است. در همین راستا تشخیص و ردیابی رویداد جهت تشخیص جدیدترین و مهم‌ترین رویدادهای جهان واقعی مورد مطالعه‌ی پژوهشگران قرار گرفته است. از اطلاعات بدست آمده از این طریق می‌توان در کاربردهای مختلف نظیر تصمیم‌گیری‌های تجاری، پیش‌بینی نتایج انتخابات و پیش‌بینی شیوع بیماری استفاده کرد. به طور کلی، تشخیص رویداد به دو دسته‌ی تشخیص رویداد گذشته‌نگر و تشخیص رویداد برخط تقسیم می‌شود. در تشخیص رویداد گذشته‌نگر، هدف یافتن رویدادها در مجموعه‌ای از اسناد از پیش جمع‌آوری شده است. درحالیکه تشخیص رویداد برخط بر شناسایی رویدادهای جدید در جریان‌های خبری برخط تمرکز دارد. در این پایان‌نامه، به تشخیص رویداد برخط در مجموعه‌ی اسناد خبری با استفاده از روش‌های تشخیص جامعه که زیر مجموعه‌ای از روش‌های خوشه‌بندی هستند، پرداخته می‌شود. روش پیشنهادی که گراف وزن‌دار عبارات انفجاری ($WebKey$) نام دارد، شبکه‌ای از عبارات انفجاری ایجاد می‌کند که این شبکه براساس هم‌رخدادی عبارات اسناد در بازه‌ی زمانی کوتاه تشکیل می‌شود. سپس دو ویژگی جدید شامل کلیک‌های کاربران بر اسناد و بر عناوین خبری از دادگان استخراج و با به کارگیری آن‌ها در الگوریتم تشخیص جامعه، خوشه‌های عبارات توصیف کننده‌ی رویداد شناسایی می‌شوند. روش پیشنهادی در مقایسه با روش‌های معروف گذشته فراخوانی را ۱۹ درصد و دقت را ۲ برابر افزایش می‌دهد.

کلیدواژه: تشخیص و ردیابی موضوع، تشخیص رویداد، وب کاوی، تشخیص جامعه

فهرست مطالب

فهرست جداول

فهرست تصاویر

فهرست اختصارات

Area-based Detection Model ..... **ADM**

Application Programming Interface ..... **API**

Breath First Search ..... **BFS**

Defense Advanced Research Projects Agency ..... **DARPA**

First Story Detection ..... **FSD**

Group Average Clustering ..... **GAC**

Hierarchical agglomerative clustering ..... **HAC**

Latent Dirichlet Allocation ..... **LDA**

Latent Semantic Indexing ..... **LSI**

Node-based Detection Model ..... **NDM**

New Event Detection ..... **NED**

Natural Language Processing ..... **NLP**

Probabilistic Latent Semantic Indexing ..... **PLSI**

Retrospective Event Detection ..... **RED**

Singular value decomposition ..... **SVD**

Support Vector Machine ..... **SVM**

Topic Detection and Tracking ..... **TDT**

Term Frequency-Inverse Document Frequency ..... **TF-IDF**

Translingual Information Detection, Extraction, and Summarization ..... **TIDES**

فصل ۱

مقدمه

۱-۱ پیشگفتار

خبر علاوه بر نقش آگاهی‌دهنده از نظر تاثیری که بر زندگی افراد جامعه می‌گذارد از اهمیت بالایی برخوردار است. از خبر می‌توان برای پیش‌بینی وضعیت بازار بورس و اقتصاد، کمک در تصمیم‌گیری‌های سیاسی، اجتماعی، اقتصادی و غیره بهره برد. اخبار از گذشته تا به امروز به روش‌های مختلفی منتشر و در دسترس عموم قرار گرفته است. با بوجود آمدن پدیده‌ی اینترنت و بستر وب، تولید و انتشار اخبار به راحتی برای همه‌ی اقشار جامعه امری آسان و راحت شد. از طرفی، رشد روزافزون اینترنت و انتشار اخبار به صورت اسناد الکترونیکی در شبکه‌های اجتماعی و انواع پایگاه‌های خبری برخط سبب شده است تا پیگیری و یافتن اطلاعات مشخص در وب، کاری زمان‌بر و دشوار شود. همچنین تشخیص تمام موضوعات به‌روز توسط یک فرد، امری غیرممکن است. بنابراین پردازش این حجم عظیم از اطلاعات نیازمند روش‌ها و ابزارهای جدید است. این مسئله، چالش‌هایی در حوزه‌ی بازیابی اطلاعات[۱] ایجاد می‌کند.

بازیابی و استخراج خبر[۲] زیرمجموعه‌ای از بازیابی اطلاعات و همچنین استخراج اطلاعات[۳] است. این دو مفهوم دو حوزه‌ی مجزا هستند. بازیابی اطلاعات به بازیابی اسناد مرتبط به پرس‌وجوی کاربر گفته می‌شود، درحالیکه استخراج اطلاعات به استخراج خودکار اطلاعات ساخت‌یافته[۴] از اطلاعات غیرساخت‌یافته[۵] یا نیمه‌ساخت‌یافته[۶] اطلاق می‌شود. بازیابی و استخراج خبر به چندین حوزه‌ی تحقیقاتی مجزا تقسیم می‌شود.

تشخیص و ردیابی موضوع (TDT[۷]) زیر مجموعه‌ای از بازیابی و استخراج اطلاعات است. انگیزه‌ی اولیه برای تحقیقات در این حوزه، فراهم کردن یک فناوری مرکزی برای ایجاد سامانه‌ای است که با نظارت بر اخبار منتشر شده، رخدادهای جالبی را که در جهان رخ می‌دهند، اعلام کند [۱]. از طرفی، شرکت‌ها و افراد به منظور توسعه‌ی الگوریتم‌های هوشمند، علاقه‌مند به استفاده از این دادگان هستند [۲]. بنابراین هدف اصلی از تشخیص و ردیابی موضوع، توسعه‌ی فناوری‌هایی است که بتوانند متون خبری را در انواع رسانه‌های خبری جست‌وجو، سازماندهی و ساختاردهی کنند [۱]. معمولاً در این حوزه از روش‌های متن کاوی[۸]، پردازش زبان

[1] Information retrieval
[2] News retrieval and extraction
[3] Information extraction
[4] Structured
[5] Unstructured
[6] Semi-structured
[7] Topic Detection and Tracking
[8] Text mining

طبیعی (NLP[1]) و تکنیک‌های خوشه‌بندی[2] استفاده می‌شود. موضوع، مفهوم کلی‌تری از رویداد است. یک موضوع می‌تواند شامل چندین رویداد باشد. رویداد به رخدادی گفته می‌شود که در زمان و مکان مشخصی روی می‌دهد. از این رو، در پژوهش‌های اخیر از عبارت تشخیص و ردیابی رویداد نیز استفاده می‌شود.

۱-۲ تاریخچه

تشخیص و ردیابی موضوع از یک مطالعه‌ی مقدماتی[3] در سال ۱۹۹۶ شروع شد. این تحقیق توسط آژانس پروژه‌های تحقیقاتی پیشرفته‌ی وزارت دفاع دولت ایالات متحده‌ی آمریکا (DARPA[4]) حمایت می‌شد و در واقع زیر مجموعه‌ای از برنامه‌ی TIDES[5] در DARPA به شمار می‌رفت [۳، ۴]. سایر مشارکت‌کنندگان در این تحقیق علاوه بر DARPA، شامل دانشگاه کارنگی ملون[6]، دراگون سیستمز[7] و دانشگاه ماساچوست[8] بودند. پژوهش فوق داده‌های خود را از رسانه‌های خبری[9] و با استفاده از سامانه‌های تبدیل گفتار به متن[10] که خبرهای تلویزیون و رادیو را به طور خودکار به متن تبدیل می‌کنند جمع‌آوری می‌کرد. درابتدا، پژوهش‌ها حول مفهوم کلی موضوع[11] صورت می‌گرفت، اما در ادامه‌ی تحقیقات درسال ۱۹۹۸ مفهوم رویداد[12] مطرح شد. هدف از تشخیص و ردیابی موضوع، شکستن متن به نقل[13] جدید است تا با نظارت بر این نقل‌ها بتواند رویدادهایی که قبلاً دیده نشده‌اند شرا ناسایی کند و این نقل‌ها را به گروه‌هایی تقسیم کند که موضوع بحث یکسانی دارند [۱]. طی این تحقیقات چندین پیکره[14] از خبرهای جمع‌آوری شده از انواع رسانه‌های خبری به زبان‌های مختلف تهیه شد که محققان در تحقیقات خود از آن‌ها استفاده می‌کردند. همچنین از این مجموعه دادگان[15] برای مسابقات TDT استفاده می‌شد. با پیدایش شبکه‌های اجتماعی[16] چالش‌ها و مفاهیم جدیدی

[1] Natural Language Processing
[2] Clustering techniques
[3] Pilot study
[4] Defense Advanced Research Projects Agency
[5] Translingual Information Detection, Extraction, and Summarization
[6] Carnegie Mellon University
[7] Dragon Systems
[8] University of Massachusetts at Amherst
[9] Newswire
[10] Speech-to-text
[11] Topic
[12] Event
[13] Story
[14] Corpus
[15] Dataset
[16] Social network

در حوزه‌ی تشخیص و ردیابی موضوع به وجود آمد. یکی از زیرمجموعه‌های جدید در این حوزه، تشخیص روندهای درحال‌ظهور[1] است که توجه بسیاری از محققان را به خود جلب کرده است. روند درحال‌ظهور، به موضوعی گفته می‌شود که در طول زمان به علاقه‌مندی و سودمندی آن افزوده می‌شود [۵].

۱-۳ کاربردها

تشخیص رویداد امکانِ فهم، استخراج و خلاصه‌سازی خودکار پیشامدهای هر رویداد در زمینه‌های مختلف از جمله بیولوژیکی، امنیت، سلامت و اقتصاد را فراهم می‌سازد. با به کارگیری روش‌های TDT، می‌توان به اطلاعات بسیاری دسترسی یافت. از جمله می‌توان به افرادی که تحت تاثیر رویداد قرار می‌گیرند، زمان و مکان رویداد، میزان خسارت ناشی از رویداد و تاثیر آن بر محیط اطراف اشاره کرد. از جمله دیگر کاربردهای تشخیص رویداد می‌توان به تشخیص زودهنگام شیوع بیماری و کشف جرایم اشاره کرد [۶]. از طرفی می‌توان در سامانه‌های استخراج اطلاعات و تصمیم‌گیری نیز از روش‌های تشخیص رویداد استفاده کرد تا کاربران به‌سرعت از رویدادهای جهان واقعی مطلع شده و در جهت بهره‌وری از شرایط به نفع خود اقدام کنند. همچنین خروجی این نوع از سامانه‌ها می‌تواند در سامانه‌های توصیه‌گر[2] نیز استفاده شود.

۱-۴ چالش‌ها

روش‌های تشخیص رویداد با چالش‌هایی روبرو هستند. از جمله می‌توان به موارد زیر اشاره کرد.

۱-۴-۱ وابستگی حوزه

روش و تکنیکی که در یک حوزه جواب می‌دهد ممکن است برای حوزه‌ی دیگر نتایج خوبی نداشته باشد و غیرقابل استفاده باشد. بنابراین بسته به شرایط و حوزه‌ی انتخابی اسناد باید از روش‌ها و تکنیک‌های متفاوتی

[1] Emerging trend detection
[2] Recommender system

جهت تشخیص رویداد استفاده کرد [۶، ۷]. با پیدایش شبکه‌های اجتماعی نیز چالش‌های بیشتری در این حوزه ایجاد شده است. با افزایش کاربران و محتوای تولید شده در این شبکه‌ها روز به روز بر اهمیت استفاده از داده‌های آن‌ها افزوده می‌شود. بنابراین نیاز به روش‌ها و ابزارهای جدید جهت استخراج اطلاعات از این داده‌ها احساس می‌شود.

۱-۴-۲ محدودیت زمانی

رویدادها باید به درستی در محدوده‌ی زمانی مشخصی تشخیص داده شوند تا بتوان از حادثه‌ی احتمالی جلوگیری کرد و یا عکس‌العمل مناسب نشان داد. بسته به حوزه‌ی انتخابی این زمان می‌تواند از ثانیه تا دقیقه تغییر کند [۶، ۷].

۱-۴-۳ صحت تشخیص

حوزه‌هایی که نیاز به تصمیم‌گیری‌های حیاتی دارند، نیازمند صحت[1] تشخیص بالایی هستند. در چنین حوزه‌هایی مانند پزشکی و بانکداری، تشخیص اشتباه منجر به خسارات مالی و یا جانی می‌شود. بنابراین انتظار می‌رود که روش‌های تشخیص رویداد از نرخ مثبت حقیقی[2] (تشخیص درست) بالا و نرخ مثبت کاذب[3] (تشخیص نادرست) پایین برخوردار باشند [۶].

۱-۴-۴ منابع داده‌ای متنوع

مجموعه‌ای عظیم از انواع داده‌ها در شبکه‌های اجتماعی و پایگاه‌های خبری برخط وجود دارد که این داده‌ها شامل اسناد متنی، تصاویر، ویدیو، صوت، داده‌های رابطه‌ای، رکوردهای چند متغیری و داده‌های فضایی و زمانی هستند. بنابراین باید داده‌ی مناسب با روش مطالعه از منابع انتخاب شود [۶].

[1] Accuracy
[2] True-positive rate
[3] False-positive rate

۱-۴-۵ انبوه داده‌ها

حجم عظیم داده‌ها نیازمند الگوریتم‌های محاسباتی با قدرت بالا و فضای ذخیره‌سازی کلان برای ذخیره، دسترسی، پالایش و پردازش داده‌ها در محدوده‌ی زمانی مشخص است [۶].

۱-۵ تعریف مساله و دستاوردها

وب جهانی یکی از غنی‌ترین منابع اطلاعاتی است و روز به روز بر تعداد کاربران و محتوای تولید شده توسط آن‌ها افزوده می‌شود. بررسی انبوهی از اطلاعاتی که هر روزه در اینترنت منتشر می‌شود در عمل کاری زمان‌بر و غیرممکن است. موتورهای جستجو یکی از بهترین گزینه‌ها برای بازیابی و پالایش اطلاعات هستند. با این حال، نیاز به سامانه‌هایی که بتوانند به صورت خودکار مهم‌ترین رویدادهای جهان واقعی را کشف کنند، احساس می‌شود. اطلاعاتی که از چنین سامانه‌هایی بدست می‌آیند می‌توانند در زمینه‌های مختلف تجاری، پزشکی، اجتماعی، سیاسی، پژوهشی و غیره کاربرد داشته باشند. بنابراین نیاز به روش‌های جدیدی برای بازیابی و استخراج اطلاعات احساس می‌شود. یکی از مطالعات مهم در این حوزه تشخیص و ردیابی رویداد است که به فرایند شناسایی رویدادهای جهان واقعی توسط تحلیل و رصد رسانه‌های جمعی به ویژه اینترنت گفته می‌شود. دقت و زمان تشخیص رویداد دو چالش جدی برای این مسئله به حساب می‌آیند.

در این پایان‌نامه با ارائه‌ی روشی مبتنی بر گراف و به‌کارگیری روش تشخیص جامعه[1] که نوعی روش خوشه‌بندی در گراف محسوب می‌شود، رویدادها در وب‌سایت‌های خبری شناسایی و خوشه‌بندی می‌شوند. روش‌های پیشین از گراف بدون وزن جهت تشخیص رویداد استفاده کرده‌اند، درحالیکه در این پژوهش از دو ویژگی جدید شامل تعداد کلیک‌های کاربران بر اسناد و عناوین خبری به همراه فراوانی عبارات سند جهت وزن‌دهی گراف و تشخیص رویداد استفاده شده است. همه‌ی این ویژگی‌ها در بازه‌ی انفجاری عبارت محاسبه شده‌اند، بدین ترتیب ویژگی زمان نیز در این روش در نظر گرفته شده است. الگوریتم‌های تشخیص جامعه عموماً زمان اجرای بالایی دارند، در این پژوهش با استفاده از روش‌های نمونه‌برداری از گراف، سازوکاری جهت بهبود سرعت اجرا به کار گرفته شده است. بدین ترتیب رویدادهای خبری در طول زمان دسته‌بندی می‌شوند و

[1] Community detection

جست‌وجو در آن‌ها برای کاربر تسهیل خواهد شد. نتایج ارزیابی نشان می‌دهد که روش پیشنهادی فراخوانی را نسبت به روش‌های پیشین ۱۹ درصد بهبود می‌بخشد و دقت را تا ۲ برابر افزایش می‌دهد.

۱-۶ ساختار کلی پایان‌نامه

ساختار کلی پایان‌نامه در ادامه بدین شرح می‌باشد. در فصل دوم، مفاهیم اولیه‌ی حوزه‌ی تشخیص و ردیابی موضوع مطرح می‌شود. در فصل سوم مروری بر پیشینه‌ی پژوهش در این حوزه خواهیم داشت. فصل چهارم به شرح روش پیشنهادی اختصاص دارد. در فصل پنجم، درباره‌ی جزییات پیاده‌سازی و ارزیابی روش پیشنهادی بحث می‌شود و فصل ششم شامل نتیجه‌گیری و ارائه‌ی پیشنهادهایی برای کارهای آینده است.