دانشکده‌ی مهندسی برق و کامپیوتر

پایان‌نامه‌ی کارشناسی ارشد در رشته‌ی مهندسی کامپیوتر- هوش مصنوعی

ارائه یک روش جدید برای بررسی میزان شباهت اسناد متنی

به وسیله‌ی

نسرین ملکوتی

استاد راهنما

دکتر علی حمزه

استادان مشاور

دکتر ستّار هاشمی

دکتر منصور ذوالقدر جهرمی

بهمن ماه ۱۳۹۱


بسم الله الرحمن الرحيم

به نام خدا

اظهارنامه

اینجانب نسرین ملکوتی دانشجوی رشته‌ی مهندسی کامپیوتر گرایش هوش مصنوعی دانشکده مهندسی اظهار می‌کنم که این پایان‌نامه حاصل پژوهش خودم بوده و در جاهایی که از منابع دیگران استفاده کرده‌ام، نشانی دقیق و مشخصات کامل آن را نوشته‌ام. همچنین اظهار می‌کنم که تحقیق و موضوع پایان‌نامه‌ام تکراری نیست و تعهد می‌نمایم که بدون مجوز دانشگاه دستاوردهای آن را منتشر ننموده و یا در اختیار غیر قرار ندهم. کلیه‌ی حقوق این اثر مطابق با آیین‌نامه‌ی مالکیت فکری و معنوی متعلق به دانشگاه شیراز است.

نسرین ملکوتی

۱۳۹۱/۱۱/۳۰

به نام خدا

ارائه یک روش جدید برای بررسی میزان شباهت اسناد متنی

به کوشش

نسرین ملکوتی

پایان‌نامه

ارائه شده به تحصیلات تکمیلی دانشگاه به عنوان بخشی از

فعالیت‌های تحصیلی لازم برای اخذ درجه‌ی کارشناسی ارشد

در رشته‌ی

مهندسی کامپیوتر (هوش مصنوعی)

از دانشگاه شیراز

شیراز

جمهوری اسلامی ایران

ارزیابی شده توسط کمیته ی پایان نامه با درجه:

دکترعلی حمزه، استادیار بخش مهندسی کامپیوتر .....

دکتر ستّار هاشمی، بخش مهندسی کامپیوتر .....

دکتر منصور ذوالقدر جهرمی، دانشیار بخش مهندسی کامپیوتر.....

بهمن ماه ۹۱

تقدیم به

مادر دلسوزم بزرگ‌ترین پشتیبانم

و

همسر مهربانم بزرگ‌ترین مشوّقم

و

استاد بزرگوارم بزرگ‌ترین مشاورم

که صبورانه یاریم کردند.

سپاسگزاری

بر خود لازم می‌دانم از زحمات استاد ارجمند جناب آقای دکتر علی حمزه که همواره در تمام مدت انجام پایان نامه مرا از راهنمایی‌ها و مساعدت‌های بی‌دریغشان بهره‌مند نمودند تشکّر کنم. همچنین از استادان مشاور خود جناب آقای دکتر ستّار هاشمی و جناب آقای دکتر منصور ذوالقدر جهرمی که با نظرات و راهنمایی‌های مفیدشان مرا در پیشبرد این پایان نامه یاری نموده‌اند کمال تشکّر را دارم.

در انتها از تمام عزیزانی که مرا در انجام این پروژه تحقیقاتی یاری نمودند کمال تشکّر و قدردانی را دارم.

چکیده

ارائه‌ی یک روش جدید برای بررسی سندهای متنی

به وسیله‌ی

نسرین ملکوتی

در سال‌های اخیر با افزایش حجم اطلاعات و داده‌های متنی، مشکلات جدیدی برای کسانی که حوزه فعالیتشان در زمینه کار با داده بود به وجود آمد. بنابراین تحقیقات بسیاری در زمینه مدل کردن اطلاعات و استخراج اطلاعات مفید از آن‌ها به عمل آمد. کاربران نیازمند ابزارهایی بودند تا با استفاده از آن‌ها به راحتی بتوانند اطلاعات مفید را از داده‌های موجود استخراج و استفاده کنند. بدین منظور مباحث بسیاری در زمینه متن کاوی و بررسی شباهت بین متون مطرح شد. برای مثال روش فضای برداری به طور گسترده‌ای در باب موضوع شباهت سنجی بین اسناد متنی سخن به عمل آورده است و مدل‌های مختلفی از معیار شباهت سنجی را معرفی کرده است. با این وجود در بسیاری مدل‌های شباهت سنجی به وجود کلمات مشترک بین اسناد متنی توجه کمتری شده است و این در حالی است که وجود کلمات مشترک بین اسناد، باعث ایجاد ابهام در روند شباهت سنجی اسناد شده و کاربران را از هدف اصلی منحرف می‌کنند.

در این پایان نامه سعی شده است یک روش برای بررسی میزان شباهت دو سند ارائه شود که با در نظر گرفتن تأثیر کلمات مشترک در بین اسناد و حذف هم‌پوشانی موجود بین اسناد متنی تخمین واقعی‌تر از میزان شباهت اسناد را به دست آورد و از این میزان شباهت برای خوشه بندی سندهای متنی[۱] استفاده شده است. این مدل شامل یک قسمت انتخاب ویژگی[۲] است که کلمات کلیدی واقع در متن را استخراج کرده، سپس با استفاده از تجزیه

[۱] Text document

[۲] Feature selection

کننده‌های متنی[۳] درخت‌های تجزیه[۴] مربوط به سندهای متنی را به دست آورده و با کمک وزن کلمات کلیدی بدست آمده از مرحله قبل، میزان شباهت بین درختان را تخمین می‌زند. برای بررسی میزان شباهت بین درختان از الگوریتم بدست آوردن تعداد زیر درختان مشابه[۵] در متن استفاده شده است. سندهای متنی به دلیل شامل بودن تعداد زیادی کلمات مشترک دارای هم‌پوشانی[۶] بسیاری هستند. وجود کلمات مشترک بین سندهای متنی ، از جمله مشکلات بررّسی دقیق میزان شباهت متن‌ها است که در صورتی که سیستم شباهت سنجی انتخاب ویژگی صحیحی از این متون به عمل آورد، قادر خواهد بود تا حدی، بر مشکل وجود هم‌پوشانی فائق آید. علاوه بر این کار کردن با متن و بدست آوردن میزان شباهت کاری زمان‌بر است، بنابراین استخراج مفهوم اصلی که از متن برداشت می‌شود از درجه اهمیت بسیاری برخوردار است.در این پایان نامه با ارائه روشی جدید برای استخراج کلمات کلیدی و با اهمیت در متن میزان شباهت بین سندهای متنی محاسبه می‌شود. در انتها با استفاده از الگوریتم‌های خوشه‌بندی[۷] از جمله الگوریتم خوشه‌بندی سلسله مراتبی[۸] و k خوشه‌بندی نزدیک‌ترین همسایه[۹] (KNN)گروه‌بندی سندهای متنی انجام شده است. نتایج آزمایشگاهی و نمودارهای مقایسه‌ای به صورت واضح نشان می‌دهند که روش پیشنهاد شده از عملکرد بالاتری نسبت به روش‌های ارائه شده مشابه دارد.

واژگان کلیدی: شباهت دو سند، انتخاب ویژگی، خوشه بندی، خوشه‌بندی سلسله مراتبی، خوشه‌بندی k نزدیک‌ترین همسایه(KNN).

[۳] Text parser
[۴] Parse tree
[۵] Isomorphm subgraph
[۶] Overlap
[۷] Clustering
[۸] Hierarchical clustering
[۹] K nearest neighbor

فهرست مطالب

عنوان **صفحه**

عنوان **صفحه**

فصل سوم: ارائه راه حل و روش‌های پیشنهادی

عنوان **صفحه**

فهرست جدول ها

عنوان **صفحه**

فهرست تصاویر

عنوان **صفحه**

عنوان صفحه

فصل اول

مقدمه و تعاریف اولیه

در این بخش یک شمایی از کار تحقیقاتی و مشکلات آن ذکر شده است. بخش ۱.۱ به معرفی سیستم‌های بازیابی اطلاعات و علت به وجود آمدن سیستم‌های بازیابی اطلاعات پرداخته است. بعضی از مفاهیم مهم و اساس کار در بخش ۱.۲ معرفی شده‌اند. در بخش ۱.۳ خلاصه‌ای از کارهای لازم که برای پیش پردازش متون انجام می‌شود ارائه شده است و در بخش ۱.۴ هدف از انجام این پایان‌نامه و چالش‌های ذکر شده است.

۱-۱ مقدمه

یکی از مشکلات بزرگ بشر در طول حیات خویش ناتوانی در دریافت اطلاعات و کم دانشی وی بوده است و هرگاه نیز اطلاعات و تجربیاتی کسب کرده است در نحوه انتقال و تبادل اطلاعات دچار مشکل شده است. این موضوع به دلیل پیشرفت‌هایی که در سال‌های اخیر حاصل شده، تبدیل به یک چالش بزرگ شده است. با پیشرفت علم و افزایش اطلاعات حجم زیادی از اطلاعات تولید شده که نیاز به کنترل و دسته بندی داشت. علاوه بر این با افزایش حجم اطلاعات، کار برای کسانی که با داده و اطلاعات سروکار داشتند مشکل شده و سبب به وجود آمدن مشکلاتی از جمله مشکلات روانی، اجتماعی، اقتصادی و غیره برای کسانی که با اطلاعات کار می‌کردند شد. با به وجود آمدن شبکه جهانی اینترنت شمار سندها به-خصوص سندهای متنی روز به روز بیشتر و بیشتر شد[۱]. انقلاب تکنولوژی در چند دهه اخیر، اطلاعات ساده و ایستا را به شکل پیچیده و دشوار و پویا درآورد. این تحول بزرگ منجر به تولید انواع اسناد و تبدیل و انتقال آن‌ها به انواع روش‌های دیگر گردیده است. انواع مختلف اسناد از جمله اسناد متنی، اسناد صوتی، اسناد تصویری و غیره که هر کدام شامل انواع متنوعی از داده‌ها می‌باشند باعث تولید چالش‌های فراوانی برای انواع عملیات روی داده می‌شوند. این امر، نیاز به سیستم‌های دقیق، سریع و مطمئن را بیش از پیش نشان داد. لذا همواره نیاز به روش‌هایی است که بتوانند انسجام ساختار را حفظ نموده و علاوه بر این توسعه و تسهیل استفاده از منابع را برای گسترش و انتخاب داده‌ای فراهم کنند[۲]،[۳]. اختراع رایانه قدمی بزرگ در مدیریت و سازماندهی اسناد بود. با استفاده از این ماشینِ اختراع بشر مدیریت داده‌ها و اطلاعات کاغذی روز به روز ماشینی‌تر و خودکارتر شدند. با پیشرفته شدن رایانه‌ها و به وجود آمدن واحدهای نرم‌افزاری و سخت‌افزاری مکانیزم‌های جستجو، بازیافت و تبدیل اطلاعات و داده‌ها تسهیل یافت. سیستم‌های مدیریت پایگاه داده که در جهت ذخیره و پردازش داده‌های ساختار یافته توسعه یافتند، تسهیل در

امر پردازش اطلاعات را بیش از پیش مهیا کردند. در داده‌های متنی به دلیل ویژگی غیر ساختار یافته آن‌ها و پیچیدگی‌های جستجو و مکانیزم‌های پردازش نیاز به نوع خاصی از تکنولوژی پردازش داده که با عنوان سیستم‌های بازیابی متون شناخته شده‌اند، احساس شد. سیستم‌های بازیافت اطلاعات در دهه‌های اخیر برای رویایی با حجم عظیمی از اطلاعات از توجه فراوانی برخوردار شدند. این امر، منجر به فراهم ساختن دسترسی به مقادیر عظیمی از داده و برآورده کردن نیازهای تبدیل، جستجو و سایر تکنولوژی‌های پردازش روی داده و تولید اطلاعات مفید و دانش شد[۴].

کار سیستم‌های بازیافت اطلاعات استخراج الگوهای مناسب از منابع عظیم داده‌ای است که اغلب در پایگاه‌های داده‌ای بزرگ ذخیره شده‌اند. اغلب مطالعات روی سیستم‌های بازیافت اطلاعات روی داده‌های ساختار یافته تمرکز داشته‌اند؛ در حالی که در واقعیت بخش عظیمی از زیرمجموعه اطلاعات در دسترس، در پایگاه‌های داده متنی ذخیره شده‌اند که شامل مجموعه‌های عظیم سندهایی از منابع گوناگون همچون عناوین اخبار، مقالات تحقیقاتی، کتاب‌ها، کتابخانه‌های عددی، پیام‌های الکترونیکی و صفحات وب هستند[۶]. امروزه اغلب اطلاعات در حوزه‌های دولتی، صنعتی، شغلی و غیره به صورت الکتریکی و در پایگاه‌های داده متنی ذخیره شده‌اند. همچنین اغلب داده‌های ذخیره شده در پایگاه‌های داده متنی به صورت نیمه ساختار یافته هستند. این بدین معنی است که داده‌های متنی نه به صورت کامل ساختار نیافته و نه به صورت کامل ساختار یافته هستند. به عنوان مثال، یک سند ممکن است شامل تعداد کمی حوزه‌های ساختار یافته مانند عنوان، نویسنده، گروه، تاریخ انتشار و غیره باشد؛ اما در عین حال شامل شمار بسیاری از اجزای متنی ساختارنیافته همانند چکیده یا متن باشد. بنابراین در حوزه‌های تحقیقاتی امروزه، مطالعات بسیاری در حیطه مدل کردن و پیاده سازی داده‌های نیمه ساختاریافته صورت گرفته است[۵]. با وجود این بدون کسب شناخت لازم از یک سند یافتن مدل مناسب و موثر جهت پردازش و استخراج مفید داده میسر نمی‌باشد. علاوه بر این، کاربران نیازمند ابزاری هستند تا سندهای پیچیده را با هم مقایسه کنند تا بتوانند آن‌ها را از نظر درجه اهمیت و ارتباطشان طبقه بندی کنند و یا اقدام به پیدا کردن الگوهایی از مدل‌های مختلف سندها و یا ترکیب کردن انواع سند با یکدیگر کنند.

اکثر سیستم‌های بازیابی اطلاعات سنتی، اطلاعات مربوط به فهرست داده‌ها را ذخیره و پردازش می‌کردند. در روش‌های سنتی، اساس مکانیزم جستجو روی کلیدهای فهرست و انواع نشانه‌های ذخیره شده برای بازیابی منابع اطلاعاتی و یا از طریق آدرس محل ذخیره شده اطلاعات اصلی استوار بود. روش‌های سنتی بازیابی اطلاعات حتی هنگامی که برای یافتن شباهت، مفاهیم را نیز علاوه بر کلمات و عبارات دخیل می‌کنند، به دلیل وجود عدم ارتباط مستقیم میان مفاهیم با معانی، ممکن است صحت شباهت بین انواع سند مورد تائید نباشد.

۱-۲ مفاهیم و اصطلاحات

در ادامه تعاریف و مفاهیم در ارتباط با کارهای انجام شده آورده شده است که در دریافت خطی مشی کلی تحقیق کمک می‌کند.

۱-۲-۱ داده کاوی[۱۰]

داده کاوی به استخراج دانش از منابع عظیم داده، اشاره دارد [۷]. در سال‌های اخیر، داده کاوی سهم عظیمی از توجه اطلاعات صنعتی و اجتماعی را به خود اختصاص داده است. مقادیر چشمگیر داده‌های قابل دسترس و نیاز فوری به تغییر داده‌های موجود و تبدیل آن‌ها به اطلاعات مفید و دانش، سرچشمه به وجود آمدن مفهوم داده کاوی شده است. بسیاری از اصطلاحات دیگر حاوی معانی مشابه با داده کاوی می‌باشند، از جمله می‌توان به استخراج دانش از داده[۱۱]، استخراج دانش[۱۲]، آنالیز داده/ الگو[۱۳]، داده شناسی[۱۴] و داده برداری[۱۵] اشاره کرد. داده کاوی می‌تواند به صورت نتیجه‌ای از تحول طبیعی و سیر تکاملی از تکنولوژی اطلاعات باشد و برای انواع مختلفی از داده حتی داده‌های زودگذر[۱۶] مانند جریان‌های

[۱۰] Data mining
[۱۱] Knowledge mining from data
[۱۲] Knowledge extraction
[۱۳] Data/pattern analysis
[۱۴] Data archaeology
[۱۵] Data dredging
[۱۶] Transient data

داده[۱۷] قابل کاربرد است. علاوه بر این، عملیاتی همچون گروه‌بندی داده‌ها بر اساس ویژگی‌های تعیین شده در مدل‌های داده‌ای، خوشه‌بندی داده‌ها بر اساس معیار شباهت تعریف شده در بین اعضای خوشه‌های داده و بازیابی داده بر اساس نوع درخواست مطرح شده، یادگیری و کشف قوانین و روابط جدید میان الگوهای داده‌ای، برخی از قابلیت‌های عملیاتی و کاربردی داده‌کاوی می‌باشد و هدف از آن، یافتن الگوی داده‌ای، ارتباطات داده‌ای و دریافت قوانین موجود میان داده‌ها می‌باشد. اطلاعات و دانش به دست آمده در حوزه‌های مختلف و همچنین کاربرد تحلیل تجاری، تشخیص عیب و حفظ حقوق مشتری برای ایجاد کنترل، کشف قوانین مربوط به داد و ستد تجاری، روابط بین مخاطبان کالا و یا کشف جرایم و تشخیص سوء استفاده از اطلاعات در بسیاری از زمینه‌ها تنها بخشی از توانمندی-های کاربردی داده کاوی می‌باشد که در سال‌های اخیر مورد توجه و تحقیق وسیعی قرار گرفته است.

۱-۲-۲ متن

متن رشته‌ای از کاراکترها و کلمات؛ مانند کلمات توصیفی، پرسشی، نمادها ، اعداد و غیره می‌باشد. کلمات موجود در متن، کلمات سخت و آسانی هستند که در قالب جمله یا پاراگراف ظاهر می‌شوند. پیام‌ها، داستان‌ها، تلگراف‌ها، پیام‌های الکترونیکی و غیره از جمله ورودی‌های ذخیره شده در قالب متن می‌باشند که با پیروی از قوانین زبان‌های طبیعی مانند فارسی ، انگلیسی، فرانسه و غیره، قابل خواندن توسط ماشین هستند.

۱-۲-۳ سند

سندها فایل‌های قابل خواندن توسط ماشین می‌باشند که شامل انواع مختلفی همچون متن و سایر اشکال دیگر همچون جداول، صداها، تصاویر ویدیویی و غیره هستند. بر اساس تعریف ارائه شده، یک سند به هر شیء حاوی اطلاعات که قابل پردازش است، اطلاق می-گردد. این اشیاء از قبیل متن، ویدیو، گرافیک، جداول و یا هر نوع دیگر از فایل که ممکن است شامل تعدادی زیرساختار دیگر مانند بخش‌ها، قسمت‌ها، پاراگراف‌ها و غیره باشند. با دسته‌بندی اسناد در گروه‌های مجزا بر اساس شباهت در محتوا یا نوع سند به منظور اهداف

[۱۷] Data streams