نشریه علمی «نوآوری‌های فناوری اطلاعات و ارتباطات کاربردی»
سال اول، شماره ۲، زمستان ۱۳۹۸؛ ص ۲۶- ۱۷

ارائه روش ترکیبی تحلیل پایگاه داده‌های خبری با استفاده از RapidMine
مطالعه موردی: متون خبری فارسی

سعید احمدیان[1*]، کریم صابری[2]،

۱- کارشناسی ارشد دانشگاه امام حسین (ع)، ۲- کارشناسی ارشد دانشگاه خواجه‌نصیرالدین طوسی

(دریافت: ۱۳۹۸/۱۱/۰۲، پذیرش: ۱۳۹۸/۱۲/۱۳)

چکیده

یکی از مهم‌ترین و پرکاربردترین رویکردها در مدیریت سامانه‌های تحت وب، استفاده از خبرگزاری‌های آنلاین و پایگاه داده‌های خبری هست. خبرگزاری‌های آنلاین و تحت وب به‌مانند قلب تپنده یک جامعه، جزء حیاتی‌ترین منابع اطلاعاتی یک جامعه به‌حساب می‌آیند. اهمیت این موضوع در مورد اخبار سازمان‌های نظامی دوچندان خواهد بود؛ بنابراین، داشتن روشی منحصربه‌فرد، صحیح و مفهومی جهت تحلیل و دسته‌بندی این منابع، می‌تواند مزایای متعددی فراهم نماید و به تصمیم گیران سازمان‌ها به‌ویژه سازمان‌های نظامی در انجام تصمیمات کمک کند. در این پژوهش، در مرحله اول به بررسی و شناخت مفاهیم انواع روش‌های داده‌کاوی می‌پردازیم. در مرحله دوم انواع روش‌های پردازش متن، از منظرهای مختلف باید صورت گیرد تا بتوان راه‌کارهای مثبت و نقاط قوت انواع روش‌ها را شناسایی کرده و برای تحلیل پایگاه داده‌های خبر بهترین روش را شناسایی کرد. در انتها پایگاه داده‌های خبر را مورد تحلیل قرار داده تا بتوان اطلاعات پنهان موجود در این داده‌ها را کشف و مورداستفاده قرار داد.

کلیدواژه‌ها:

داده‌کاوی، متن‌کاوی، پایگاه داده‌های خبر

۱- مقدمه

اخبار موجود در دنیای امروز غالباً به‌عنوان یک پایگاه داده ثبت اخبار و به‌صورت آرشیوی انجام می‌شود درحالی‌که در میان آن‌ها دانش‌ها و روابط پنهان بسیار مفیدی قرار دارند. کشف این دانش‌ها می‌تواند برای تصمیم گیران و تحلیل گران در حوزه‌های مختلف و به‌ویژه در مسائل مرتبط با پدافند سایبری و تحلیل خبرها به‌صورت آنی بسیار کاربردی و جالب‌توجه باشد. با توجه به حجم عظیم این داده‌ها بایستی با استفاده از روش‌ها و فن‌های علمی، این داده‌ها تجزیه‌وتحلیل شوند و روابط و دانش‌های پنهان میان آن‌ها استخراج و در دسترس مدیران و تصمیم گیران قرار گیرند. با این کار بازدهی فرایند تحلیل و شناسایی خبرها به‌صورت هدفمند افزایش می‌یابد، به‌این‌علت که تحلیل خبرها از حالت تحلیل فرد کاربر خارج‌شده و توسط روش‌های داده‌کاوی با بازدهی بسیار بالاتری نسبت به تحلیل دستی کاربر، به انجام می‌رسد. استفاده از این روش‌ها می‌تواند به کلیه سازمان‌های داخل کشور اعم از نظامی و غیرنظامی این توانایی را بدهد که اخبار رسیده به سازمان متناسب با الگوی تعیین‌شده همان سازمان، تحلیل شود و در اختیار تصمیم گیران آن سازمان قرار گیرد. تصمیم‌گیران بدین‌وسیله توانایی تصمیماتی صحیح، مناسب با موقعیت موجود، سریع و با دانش بیشتری پیدا خواهند نمود.

لذا در این تحقیق با تشریح دقیق کاربرد روش‌های داده‌کاوی و متن‌کاوی در تحلیل خبرهای منتشرشده و کشف روابط پنهان میان آن‌ها، چگونگی استخراج روابط و نتایج کارامل در سامانه پایگاه خبری ارائه خواهد شد. درنهایت با دانش‌های پنهان در پایگاه داده در دسترس قرارگرفته که تحلیل و تفسیر خواهند شد، یک چارچوب برای تحلیل خبرها و چگونگی روابط پنهان میان خبرها بر اساس روش‌های پیش بینانه داده‌کاوی ارائه خواهد شد. به‌طورکلی اهدافی که از ارائه این روش دنبال می‌شود را می‌توان بدین گونه ذکر کرد:

۱. تعیین کلمات کلیدی و ماهیت خبرها بر اساس الگوهای متن‌کاوی و تهیه گزارش‌ها تحلیلی از آن‌ها.
۲. کشف دانش‌های پنهان از خبرهای تولیدشده بر اساس روند سری‌های زمانی موجود در بین آن‌ها.
۳. کشف و تحلیل روابط پنهان میان کلمات کلیدی و ماهیت خبرها در گذشته.
۴. کشف کلمات مهم در خبرهای منتشرشده که تأثیرگذاری بسیاری در ذهنیت مردم و خوانندگان خبرها دارند و تعیین جهت‌گیری‌های ماهیت خبرها بر اساس روش‌های داده‌کاوی.

[1] رایانامه نویسنده پاسخگو: s.ahmadian@ihu.ac.ir

۵. دسته‌بندی و خوشه‌بندی صحیح اطلاعات و داده‌ها که موجب جلوگیری از نشر اخبار محرمانه نظامی و غیرنظامی خواهد شد.

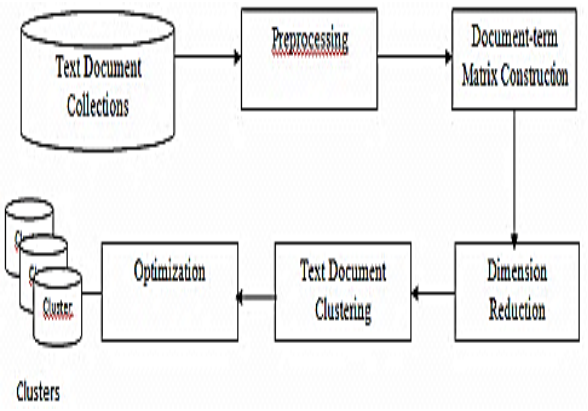
۲- تعریف مفاهیم

متن‌کاوی: استخراج اطلاعات از متن شامل ارائه یک قالب ساختارمند مانند یک پایگاه داده از اطلاعات دلخواه موجود در متن هست [۱۴]. متن‌کاوی به مفهوم جستجوی الگوها در متون بدون ساختار است. متن‌کاوی ترکیبی از فن‌های کاوش داده و استفاده آن‌ها با کمک لغت‌شناسی موضوعی هست [۱۵].

دو معیار اصلی برای ارزیابی کارایی سامانه‌های اطلاعات وجود دارد. اول این‌که، چه درصدی از اطلاعات استخراجی صحیح هستند و دوم این‌که، چه درصدی از اطلاعات موجود در متن استخراج یافته‌اند [۱۷].

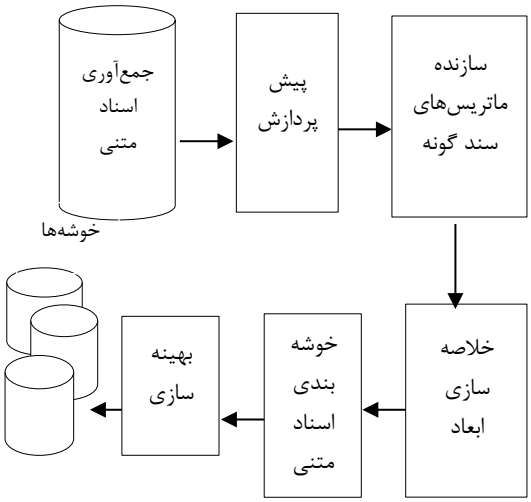
یکی از مراحل متداول در تحلیل اسناد متنی خوشه‌بندی این اسناد هست. با خوشه‌بندی، اسناد مشابه را می‌توان با سرعت کاوش کرد، به‌آسانی می‌توان موضوعات و زیر موضوعات آن‌ها را درک کرد و به‌طور مؤثری بین آن‌ها برای بسیاری از کاوش‌ها پرس‌وجو انجام داد. خوشه‌بندی اسناد متنی، نقش مهمی را در سازمان‌دهی مؤثر سند، خلاصه‌سازی، استخراج موضوع و بازیابی اطلاعات ایفا می‌کند و یک ابزار کارا برای مدیریت سربار اطلاعات است [۱۶].

می‌توان متن‌کاوی را گونه‌ای اختصاصی از خوشه‌بندی دانست با این تفاوت که به دلیل داشتن ابعاد بالاتر کمی دشوارتر هست. قدم‌های اساسی برای متن‌کاوی در شکل (۱) به نمایش درآمده است.



شکل (۱): مراحل متن کاوی .

۳- تعاریف متن‌کاوی



شکل (۲): مراحل متن کاوی.

می‌توان گفت که متن‌کاوی از تکنیک‌های بازیابی اطلاعات، استخراج اطلاعات همچنین پردازش کردن زبان طبیعی[1] استفاده می‌کند و آن‌ها را به الگوریتم‌ها و متدهای KDD، داده‌کاوی، یادگیری ماشین و آماری مرتبط می‌کنند. با توجه به ناحیه‌های تحقیق گوناگون، بر هر یک از آن‌ها می‌توان تعاریف مختلفی از متن‌کاوی در نظر گرفت در ادامه برخی از این تعاریف بیان می‌شوند:

- *متن‌کاوی = استخراج اطلاعات:* در این تعریف متن‌کاوی متناظر با استخراج اطلاعات در نظر گرفته می‌شود (استخراج واقعیت‌ها[2] از متن).
- *متن‌کاوی = کشف داده متنی:* متن‌کاوی را می‌توان به‌عنوان متدها و الگوریتم‌هایی از فیلدهای یادگیری ماشین و آماری برای متن‌ها باهدف پیدا کردن الگوهای مفید در نظر گرفت. برای این هدف پیش‌پردازش کردن متون ضروری است. در بسیاری از روش‌ها، متدهای استخراج اطلاعات، پردازش کردن زبان طبیعی یا برخی پیش‌پردازش‌های ساده برای استخراج داده از متون استفاده می‌شود. سپس می‌توان الگوریتم‌های داده‌کاوی را بر روی داده‌های استخراج‌شده اعمال کرد.
- *متن‌کاوی = فرایند* KDD: که در بخش بعد به‌طور کامل توضیح داده شده است.

[1] Natural language processing (NLP)
[2] Facts

۴- کشف دانش و ارتباط آن با متن‌کاوی

کشف دانش در پایگاه داده‌ها (KDD)، این لغت به همه شیوه‌هایی اشاره دارد که هدف آن‌ها پی‌بردن به ارتباط و نظم بین اطلاعات قابل‌مشاهده است. لغت KDD برای توصیف همه مراحل استخراج اطلاعات از پایگاه داده و نیز بیان اهداف کارهای اولیه کاربرد قوانین تصمیم‌گیری است. در کل KDD فرآیند یافتن اطلاعات و الگوهای مفید از داده را گویند و داده‌کاوی بهره‌گیری از الگوریتم‌هایی برای یافتن اطلاعات مفید در فرآیند KDD است.

ازجمله خصوصیاتی که برای اندازه‌گیری کیفیت الگوهای پیداشده در داده می‌توان استفاده کرد عبارت‌اند از: قابلیت فهم انسان، اعتبارسنجی با معیارهای آماری و تازگی و مفید بودن. کشف دانش در پایگاه داده را می‌توان به‌عنوان یک فرایند که به‌وسیله چندین گام پردازش کردن[1] تعریف می‌شود، در نظر گرفت. این گام‌ها به‌منظور استخراج الگوهای مفید باید بر روی مجموعه داده‌ای اعمال شوند. این گام‌ها به‌صورت تکراری اجرا می‌شوند و برخی گام‌ها نیاز به بازخورد از کاربر دارند. یک کاربر سامانه KDD به‌منظور انتخاب زیرمجموعه صحیحی از داده‌ها باید درک بالایی از قلمرو داده‌ها، رده مناسبی از الگوها و معیار خوبی برای الگوهای جالب داشته باشد؛ بنابراین سامانه KDD باید ابزارهایی با اثر تعاملی داشته باشد نه سامانه‌های تجزیه‌وتحلیل خودکار. طبق [۱۹] گام‌ها را می‌توان به‌صورت زیر بیان کرد: ۱) درک کردن کسب‌وکار ۲) درک کردن داده ۳) آماده‌سازی داده ۴) مدل کردن ۵) ارزیابی ۶) deployment. مرحله پیش‌پردازش غالباً یکی از مراحل زمان‌بر و درعین‌حال بسیار مهم در کسب نتیجه مطلوب است. مخصوصاً در متن‌کاوی که نیاز به متدهای پیش‌پردازش کردن خاصی برای تبدیل داده متنی به فرمتی که برای الگوریتم‌های داده‌کاوی مناسب است، داریم.

۵- ناحیه‌های جستجوی مرتبط

سه روش اساسی در مواجهه با این حجم وسیع از اطلاعات غیر ساخت‌یافته وجود دارد که عبارت‌اند از: بازیابی اطلاعات[2]، استخراج اطلاعات[3] و پردازش زبان طبیعی

بازیابی اطلاعات: اصولاً مرتبط است با بازیابی مستندات و مدارک. کار معمول در بازیابی اطلاعات این است که با توجه به نیاز مطرح‌شده از سوی کاربر، مرتبط‌ترین متون و مستندات و یا درواقع کلمه را از میان دیگر مستندات یک مجموعه بیرون بکشد. این عمل یافتن دانش نیست بلکه تنها آن مجموعه‌ای از کلمات را که به نظرش مرتبط‌تر به نیاز اطلاعاتی جستجوگر است را به او تحویل می‌دهد. این روش به‌واقع هیچ‌دانشی و حتی هیچ اطلاعاتی را برایمان به ارمغان نمی‌آورد.

پردازش زبان طبیعی: هدف کلی NLP[4] رسیدن به یک درک بهتر از زبان طبیعی توسط کامپیوترهاست. تکنیک‌های مستحکم و ساده‌ای برای پردازش کردن سریع متن به کار می‌روند. همچنین از تکنیک‌های تحلیل زبان‌شناسی نیز برای پردازش کردن متن استفاده می‌شود.

استخراج اطلاعات: هدف روش‌های استخراج اطلاعات، استخراج اطلاعات خاص از سندهای متنی است. استخراج اطلاعات می‌تواند به‌عنوان یک‌فاز پیش‌پردازش در متن‌کاوی بکار برود. استخراج اطلاعات عبارت‌اند از نگاشت کردن متن‌های زبان طبیعی (مثلاً گزارش‌ها، مقالات journal، روزنامه‌ها، ایمیل‌ها، صفحات وب، هر پایگاه داده متنی و ...) به یک نمایش ساخت‌یافته و از پیش تعریف‌شده یا قالب‌هایی که وقتی پر می‌شوند، منتخبی از اطلاعات کلیدی از متن اصلی را نشان می‌دهند. یک‌بار اطلاعات استخراج‌شده و سپس اطلاعات می‌توانند در پایگاه داده برای استفاده‌های آینده، ذخیره شوند.

۶- روش‌های پیش‌پردازش کردن متون

برای کاوش کردن[5] مجموعه بزرگی از اسناد ضروری است که اسناد پیش‌پردازش شوند و اطلاعات در یک ساختار داده‌ای مناسب برای پردازش‌های بعدی ذخیره شوند. روش‌های اصلی پیش‌پردازش متون عبارت‌اند از: مدل فضای بردار[6] [۲۲]، مدل احتمالی [۲۳] و مدل منطقی [۲۴].

۷- روش‌های متن‌کاوی

دلیل اصلی به کار بردن روش‌های داده‌کاوی برای اسناد متنی، ساختاربندی کردن آن‌هاست. ساختارهای معروف عبارت‌اند از: کاتالوگ‌های کتابخانه یا نمایه‌های کتاب. مشکل نمایه‌های طراحی‌شده به‌صورت دستی، زمان موردنیاز برای نگه‌داری آن‌ها است؛ بنابراین برای منابع اطلاعاتی که خیلی تغییر می‌کنند مثل

[1] Processing
[2] Information Retrieval
[3] Information Extraction
[4] Natural language processing
[5] Mining
[6] Vector space

وب مناسب نیستند. متدهای موجود برای ساختاربندی کردن مجموعه‌ها عبارت‌انداز: روش‌های رده‌بندی و روش‌های خوشه‌بندی.

۷-۱- رده‌بندی

هدف از رده‌بندی متون نسبت دادن کلاس‌های از پیش تعریف‌شده به اسناد متنی است؛ مثلاً یک خبر جدید که وارد می‌شود بگوییم متعلق به کلاس ورزشی یا سیاسی یا هنری. برای رده‌بندی اسناد روش‌های گوناگونی به کار می‌روند. در رده‌بندی یک مجموعه آموزش از اسناد وجود دارد که برای این مجموعه کلاس‌ها مشخص است. با استفاده از این مجموعه مدل رده‌بندی مشخص می‌شود، سپس با استفاده از آن کلاس سند جدید که وارد می‌شود، مشخص می‌گردد. برای اندازه‌گیری کارایی مدل رده‌بندی، یک مجموعه آزمودن که مستقل از مجموعه آموزش است در نظر گرفته می‌شود؛ و برچسب‌هایی که برای این اسناد توسط مدل تخمین زده می‌شود با برچسب واقعی اسناد مقایسه می‌شود. روش‌های رده‌بندی:

- انتخاب‌ترم ایندکس[1]
- رده‌بندی کننده Naive Bayes
- رده‌بندی نزدیک‌ترین همسایه
- درخت تصمیم‌گیری

۷-۲- الگوریتم خوشه‌بندی

انتخاب الگوریتم بیشتر وابسته به مجموعه داده و وظیفه‌ای است که باید انجام شود. خوشه‌بندی سلسله مراتبی و خوشه‌بندی رابطه‌ای باینری از معروف‌ترین الگوریتم‌های خوشه‌بندی هستند. در روش‌های قبلی، خوشه‌ها در درخت‌های خوشه‌ای (سلسله مراتبی) قرار می‌گرفتند. به‌طوری‌که خوشه‌های مرتبط در یک شاخه از درخت بودند.

این تحلیل به‌طور خودکار می‌تواند برای وظایف زیادی مفید واقع شود. همچنین می‌تواند یک دیدی از محتویات یک مجموعه بزرگ اسناد فراهم کند. وظیفه دیگر، شناسایی ساختارهای پنهان در گروه‌های اشیاء است. فرآیند پیدا کردن اطلاعات مربوط می‌تواند آسان و دقیق شود. به‌علاوه، تمایلات جدید (سیاست‌های جدید) که بیان‌نشده‌اند در اسناد دیگر، می‌تواند در اسناد کشف گردد و سرانجام کپی اسناد در یک مجموعه می‌تواند حذف شود.

تحلیل خوشه‌بندی، یک فن آماری قدیمی است. اشیاء بر اساس معیار شباهت و فاصله عددی بین اشیاء به گروه‌هایی تقسیم می‌شوند؛ اما هر نوع شباهتی را نمی‌توان به‌صورت عددی تخمین زد. برای مثال فاصله بین مفاهیم UMIST و OLSE. حتی اگر این فواصل بتوانند به اعداد انتقال داده شوند، ممکنه انتقال مشکل باشد؛ بنابراین، تجزیه و تحلیل خوشه مبتنی بر فاصله قدیمی در زمان پردازش کردن مفاهیم ممکنه نتایج بامعنی تولید نکند.

در خوشه‌بندی مفهومی، خوشه‌ها تنها مجموعه‌ای از اشیاء با شباهت عددی نیستند. خوشه‌ها به‌عنوان گروهی از اشیا که با یکدیگر یک مفهوم را نشان می‌دهند، هستند. برای خوشه‌بندی مفهومی ما به مجموعه‌ای از صفات برخی اشیاء نیاز داریم (یک‌زبان توصیف برای مشخص کردن خوشه‌های چنین اشیایی) و یک معیار کیفیت خوشه‌بندی است. هدف، تقسیم‌بندی کردن اشیاء در خوشه‌ها به‌گونه‌ای است که معیار کیفیت بیشینه شود و درعین‌حال تعیین کردن توصیفات عمومی از این خوشه‌هاست. توجه شود که در خوشه‌های مفهومی خصوصیات خوشه با بررسی و دقت به فرآیند مشخص کردن خوشه‌ها به وجود می‌آید. این‌یک تفاوت اصلی بین خوشه‌بندی مفهومی و قدیمی است. درروش خوشه‌بندی قدیمی خوشه‌ها مطابق با یک معیار شباهت مشخص می‌شوند. این معیار شباهت یک تابع است که فقط خصوصیات اشیاء مقایسه می‌شوند. در مقابل در خوشه‌بندی مفهومی به شرح یا توصیف هم توجه می‌شود.

۷-۳- روش‌های ترکیبی

تعداد زیادی روش در فاز استخراج دانش وجود دارد. درعین‌حال تمام این روش‌ها را شاید بتوان به دودسته اصلی تقسیم کرد. این دودسته اصلی، روش‌های مبتنی بر کارایی و روش‌های مبتنی بر دانش هستند. درروش اول، طراحان نگران کارایی سامانه هستند و به‌گونه‌ای سامانه را طراحی می‌کنند که بهترین کارایی و سرعت را داشته باشد. روش‌های رایج‌تر در این نوع نگرش، روش‌های آماری و نیز شبکه‌های عصبی هستند. روش‌های آماری بر پایه هر نوع اطلاعات آماری است که از متون قابل استخراج است. مواردی چون تکرار لغات به‌تنهایی، تکرار لغات باهم و چیزهایی شبیه آن.

در سوی دیگر روش‌های مبتنی بردانش قرار دارند که از زاویه دید دیگری به این مسئله نگاه می‌کنند. آن‌ها سعی می‌کنند اولاً تا حد ممکن مفاهیم موجود را از داخل مجموعه متون استخراج کنند و ثانیاً بین این مفاهیم روابطی برقرار کنند. استفاده از این روش بسیار وابسته به NLP است. در حقیقت این هدفی است که NLP نیز آن را دنبال می‌کند و آن درک متن است. دستگاه‌هایی

[1] Index term selection

که از این روش‌ها استفاده می‌کنند در حال حاضر زیاد نیستند ولی DR-LINK [۲۵] از دانشگاه Syracus یکی از آن‌هاست.

۷-۳-۱- روش Discotex[1]

این روش توسط Kanya در سال ۲۰۰۷ ارائه شده است [۲۰]. این روش یک چارچوب جدید برای متن‌کاوی بر اساس یکپارچه کردن سامانه استخراج اطلاعات (IE) و ماژول استنتاج کردن قوانین استاندارد (KDD) ارائه می‌کند. IE، اسناد متنی را به داده ساخت‌یافته‌تر تبدیل می‌کند. درواقع تکه‌های خاصی از داده را در سندهای به زبان طبیعی جستجو می‌کند و مجموعه‌ای از اسناد متنی نیمه ساخت‌یافته را به پایگاه داده ساخت‌یافته‌تری تبدیل می‌کند. در این روش RAPIER و BWI[2] برای ساختن IE استفاده‌شده است. سپس پایگاه داده ساخته شده به‌وسیله ماژول IE توسط ماژول KDD برای کاوش بیشتر دانش استفاده می‌شود. در یک نسخه بهبودیافته از این روش از قوانین به‌دست‌آمده از ماژول KDD برای پیش‌بینی کردن اطلاعات ازدست‌رفته و بهبود دقت ماژول IE استفاده می‌شود. برای ساختن ماژول KDD از APRIORI و RIPPER استفاده‌شده است.

۷-۳-۲- روش Textminer

در این روش ابتدا داده نیمه‌ساختیافته تغییر می‌یابد مثلاً سندها به داده ساختاریافته ذخیره‌شده در یک پایگاه داده تبدیل می‌شوند. کامپوننت دوم تکنیک‌های داده‌کاوی را روی خروجی کامپوننت اول اعمال می‌کند بیشتر روش‌ها برای متن‌کاوی، الگوریتم‌های کاوش را روی برچسب‌های نسبت داده‌شده به هر سند اعمال می‌کنند. این برچسب‌ها ممکن است کلمات کلیدی استخراج‌شده از سند یا فقط فهرستی از کلمات در سند موردنظر باشند. درروش textminer الگوریتم‌های کاوش بر روی ترم‌ها (دنباله معنی‌دار از کلمات مثلاً department of computation) ترکیب‌شده با رویدادهای (مجموعه معنی‌دار از ترم‌ها، مثلاً در یک دامنه مالی، خرید بین شرکت A و B) استخراج‌شده از سندها اعمال می‌شود. نویسندگان این مقاله معتقد هستند که مهم‌ترین فاکتورهای مشخصه که یک سند را توصیف می‌کنند ترم‌ها و رویدادهای بیان‌شده در سند هستند. این اطلاعات در یک جدول به نام EvantType نگه‌داری می‌شوند. استخراج اطلاعات فنّاوری مهمی است که یک گام پیش‌پردازش دارد. در این روش یک‌بار اطلاعات استخراج می‌شوند و سپس اطلاعات می‌توانند در پایگاه داده ذخیره شوند و برای پرس‌وجو کاوش گردند و در زبان طبیعی خلاصه شوند. اولین گام ضروری پیش‌پردازش کردن زبانی (زبان‌شناختی) است. این گام شامل تعدادی فن‌های زبان‌شناختی مثلاً tokenization، بخشی از برچسب‌گذاری گفتار و... است. اهداف کلی این مقاله را می‌توان به دودسته تقسیم نمود:۱) مدیریت کردن اطلاعات ذخیره‌شده در پایگاه داده متنی (مجموعه‌ای اسناد) ۲) استخراج کردن دانش مفید.

این روش از دو کامپوننت تحلیل متن[3] و داده‌کاوی تشکیل‌شده است. کامپوننت اول داده‌های نیمه ساخت‌یافته را به داده‌های ساخت‌یافته‌تر ذخیره‌شده در پایگاه داده تبدیل می‌کند؛ و کامپوننت دوم تکنیک‌های داده‌کاوی را بر روی خروجی کامپوننت اول اعمال می‌کند. هدف این روش مدیریت کردن اطلاعات (طبقه‌بندی کردن سندها در دسته‌های مناسب) و کاوش داده برای کشف کردن دانش مفید است؛ بنابراین در این روش ابتدا ترم‌ها و رویدادها استخراج‌شده و در پایگاه داده ذخیره می‌شوند. سپس الگوریتم خوشه‌بندی مناسبی (استفاده از الگوریتم Rock و مفهوم لنیک) بر روی پایگاه داده حاصل‌شده اعمال‌شده و اسناد گروه‌بندی می‌شوند به‌طوری‌که اسناد مشابه در یک گروه قرار گیرند. سپس یک الگوریتم رده‌بندی مناسب (درخت تصمیم‌گیری) برای معتبر سازی بیشتر نتایج حاصل از خوشه‌بندی و بهره‌برداری بهتر از دانش کشف‌شده اعمال می‌شود. برای جزییات بیشتر این روش می‌توان به برای جزییات بیشتر این روش می‌توان به مراجعه کرد.

۸- روش تحقیق و چارچوب پیشنهادی

در تحقیقاتی که امروزه صورت می‌گیرد از انواع روش‌ها استفاده می‌شود. این روش‌ها یا به‌صورت کمی و با استفاده از آمار و ارقام به نتیجه موردنظر می‌رسد، یا با استفاده از روش‌های کیفی. روش‌های تحقیق را درصورتی‌که برداری دوطرفه در نظر بگیریم، می‌توان گفت که یک‌طرف از بردار را روش‌های کمی و طرف دیگر را روش‌های کیفی تشکیل می‌دهند. روش‌های کمی با تکیه‌بر آمار و ارقام انجام می‌شود و روش‌های کیفی به کمک انواع فن‌های مشاهده و مصاحبه، به جمع‌آوری اطلاعات می‌پردازد. ویژگی مشترک این دو روش، مجهز نمودن محققین در تمام مراحل تحقیق به انواع فن‌های جمع‌آوری اطلاعات و

[1] Discovery from text extraction
[2] Boosted Wrapper induction
[3] Text Analysis Component

تجزیه‌وتحلیل‌های متنوع است. درواقع روش‌های کمی درمجموع با شمارش و اندازه‌گیری جنبه‌هایی از زندگی اجتماعی سروکار دارند، درحالی‌که روش‌های کیفی بیشتر با تولید توصیف‌های استدلالی و کشف معناها و تغییرهای کنشگران اجتماعی سروکار دارند.

بااین‌حال تمام روش‌های تحقیق، به استفاده از کمیت یا کیفیت خلاصه نمی‌شوند، بلکه می‌توان رویکردی را اتخاذ نمود که بسته به نوع تحقیق، از هردو نوع روش استفاده کند. به این نوع از روش‌ها، ترکیبی می‌گویند [۲۲- ۲۰]. درواقع ظهور و استفاده از روش‌های ترکیبی به منظر قوت دادن به تحقیقات صورت می‌گیرد [۱۸].

در تحقیق حاضر نیز از روش ترکیبی استفاده‌شده است. دلیل این امر استفاده از پایگاه داده‌های متنی متشکل از خبر است. این پایگاه داده‌ها به‌صورت کمی و کیفی، از انواع خبرها گردآوری‌شده است. همچنین به دلیل استفاده از روش‌های داده‌کاوی از استاندارد و متدولوژی CRISP برای نشان دادن چارچوب تحقیق استفاده می‌کنیم. شکل زیر چارچوب پیشنهادی تحقیق را نشان می‌دهد. به‌منظور تحلیل موردنظر در چارچوب ذکرشده، مرور ادبیات مباحثی مانند داده‌کاوی، تحلیل خبر، الگوریتم k-means و متن‌کاوی و انواع روش‌های آن به‌طور مفصل موردبررسی قرارگرفته است. این بررسی به‌منظور احاطه کامل‌تر بر مباحث و به دست آوردن ایده صورت گرفته است. فاز بعدی از تحقیق شامل بررسی سامانه موجود برای پیاده‌سازی تحلیل است. برای این منظور تحلیل‌های مختلف متن و خبر به‌عنوان مورد مطالعاتی مدنظر قرار گرفت. دلیل این انتخاب وجود داده‌های با حجم بالا در این حوزه و ارتباط تنگاتنگ با مباحث تحلیلی است.

ما در طول این تحقیق در راستای شناخت فضای موردبحث از انواع صفحات وب در زمینه اخبار و مرور ادبیات این حوزه استفاده کردیم. مرحله بعد مربوط به فاز شناخت داده و آماده‌سازی آن جهت استفاده در مرحله پیاده‌سازی است. در این مرحله بررسی می‌شود که چه داده‌هایی موردنیاز است و این داده‌های موردنیاز در چه قالبی، مناسب فاز پیاده‌سازی است. در این تحقیق از دو پایگاه داده مختلف استفاده شده است. اولین آن‌ها پایگاه داده اطلاعات نوع خبر است.

دومین پایگاه داده مربوط به متن خبرها هست که درنهایت بتوان تحلیلی دقیق‌تری به کمک الگوریتم‌های داده‌کاوی برای این بخش ارائه داد. برای اینکه بدانیم چه داده‌هایی موردنیاز است، ابتدا باید مرور ادبیات را به‌طور کامل انجام دهیم و سامانه موجود را به‌خوبی بشناسیم. پس از جمع‌آوری داده لازم است تا داده‌ها پاک‌سازی شوند. یکی از مشکلات شایع داده‌ها پایین بودن کیفیت آن‌ها است، عملیاتی که مشکل کیفیت را برطرف می‌کند، پاک‌سازی داده نامیده می‌شود [۲۵]. حذف رکوردهای پرت یکی از اقدام‌ها برای پاک‌سازی داده است. داده‌های تکراری یا دو نسخه‌ای باید شناسایی و حذف شوند. همچنین داده‌های مغشوش را نیز می‌بایست شناسایی و ترمیم گردند.

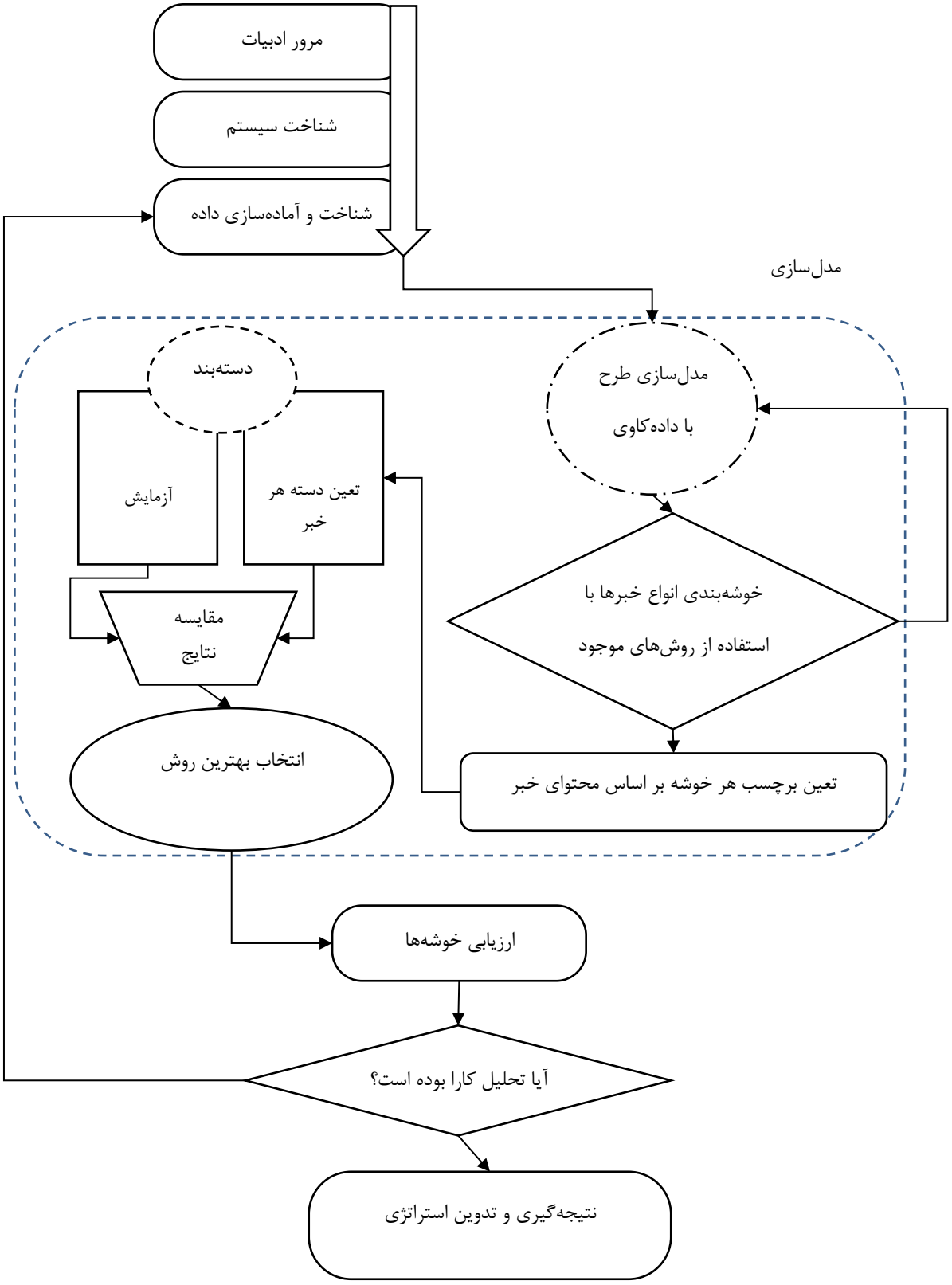
مرحله بعد مدل‌سازی است. در این مرحله دو فاز عملیاتی صورت گرفته است. فاز اول مربوط به استفاده از الگوریتم k-means جهت خوشه‌بندی انواع متن‌های خبری است بدین منظور که مشخص شود هر متن خبر مربوط به کدام حوزه خبری هست. فاز دوم تحلیل دسته‌بندی اخبار موجود به کمک روش‌هایی مانند درخت تصمیم است که ارائه‌دهنده الگوریتمی جهت مشخص کردن نوع اخبار و دسته‌بندی آن‌هاست. فاز اول یعنی استفاده از الگوریتم خوشه‌بندی k-means به کمک نرم‌افزار Rapid Miner و فاز دوم نیز به کمک نرم همین نرم‌افزار صورت گرفته است. در این مرحله به دنبال توسعه الگوریتم تحلیلی و بهبود نتایج حاصله هستیم.

۹- پیاده‌سازی طرح و نتیجه‌گیری

داده‌های موجود جهت تحلیل و پیاده‌سازی در پروژه مربوط به خبرهای مختلف از خبرگزاری‌های اینترنتی هست. پایگاه داده‌ها شامل اطلاعات جمع‌آوری‌شده از حدود ۱۰۰ پایگاه خبری اینترنتی با سه زبان مختلف فارسی، انگلیسی و عربی بود. درمجموع دو جدول مختلف در این پایگاه داده وجود دارد.

یک جدول مربوط به اطلاعات خبر منتشره که حاوی اطلاعاتی مانند شناسه گروه، شناسه منبع، شناسه گزارش گیر، شناسه نوع، شناسه خبر، آدرس خبر، تاریخ انتشار، تاریخ ایجاد، شمارنده فایل، مسیر فایل، آدرس فایل و وضعیت خبر هست. جدول دوم با نام جدول متن نیز شامل اطلاعاتی مانند شناسه متن، شناسه خبر، شناسه فایل، شناسه زبان، متن و نوع متن می‌باشد. تعداد کل داده‌ها در جدول خبر شامل ۱۶۳۴۸۶ سطر یکتا و در جدول متن ۴۴۷۴۰۷ سطر یکتا هست. دلیل تمایز بین تعداد موجود در دو جدول نوع داده‌های[1] مختلف در جدول متن هست، به این صورت که هر شناسه خبر یکتا در جدول خبر، انواع مختلفی متن ذخیره‌شده در جدول متن را داراست، مانند عنوان خبر، متن خبر و خلاصه خبر مربوط به یک خبر خاص.

[1] Type ID



شکل (۳): روش انجام تحقیق

برای شناسایی خبرهای معتبر بین دو جدول از عملیات پیوند[1] استفاده کردیم که تعداد سطرهای حاصله از این عملیات ۴۴۷۳۹۲ است. به این معنی که سطرهای مختلف بین دو جدول شناسایی‌شده و تعداد کل به دست می‌آید. با انجام این عمل مشخص می‌شود که تعدادی داده اضافی در هر دو جدول وجود دارد. تعداد شناسه خبرهایی که در جدول خبر هستند ولی در جدول متن پیدا نشده‌اند برابر ۳۳۶۷ بود، به این معنی که شناسه خبر موجود ولی متنی برای خبر وجود ندارد. همچنین تعداد خبرهایی که در جدول خبر هستند ولی در جدول (۱) متن به آن‌ها ارجاعی داده نشده ۱۵ عدد است.

همان‌گونه که در بالا بیان شد در جدول متن ۱۵ رکورد وجود داشتند که شناسه خبر آن‌ها در جدول خبر یافت نشد. این ۱۵ رکورد درواقع ۶ شناسه خبر متمایز بودند؛ بنابراین، بایستی این شناسه خبرها در جدول متن و درنهایت جدول پیوند حذف گردند.

```
delete
from dbo.News_Text
where NewsId in (select t2.NewsId
                 from dbo.News_Text t2
                 where t2.NewsId not in (select t1.NewsId
                                         from dbo.News_News t1))
```

الگوریتم (۱): حذف خبری که شناسه خبر آن‌ها در جدول متن یافت نشد

در جدول خبر ۳۳۶۷ رکورد وجود داشتند که شناسه خبر آن‌ها در جدول متن یافت نشد؛ بنابراین بایستی این شناسه خبرها در جدول خبر و نهایتاً پیوند حذف گردند.

```
delete
from dbo.News_News
where NewsId in (select t1.NewsId
                 from dbo.News_News t1
                 where t1.NewsId not in (select t2.NewsId
                                         from dbo.News_Text t2))
```

الگوریتم (۲): حذف خبری که شناسه خبر آن‌ها در جدول متن یافت نشد

با توجه به این‌که هدف اصلی پروژه متن‌کاوی اطلاعات موجود از داده‌های خبری فارسی حاصل از خبرگزاری‌های جمهوری اسلامی ایران است، باید کلیه رکوردهای غیرفارسی از جدول متن پاک گردد. در این راستا ۷۴۶۲ رکورد و متن انگلیسی و همچنین ۲۹۴۸ رکورد و متن عربی شناسایی و حذف گردید.

```
DELETE
FROM DBO.News_Text
WHERE LanguageId = '2'
```

الگوریتم (۳): واکشی خبر بر اساس نوع زبان

```
DELETE
FROM News_Text
WHERE LanguageId = '3'
```

الگوریتم (۴): واکشی خبر بر اساس نوع زبان

ستون «شناسه زبان» که نشان‌دهنده نوع زبان خبر بود، حذف گردید. این ستون شامل مقادیر متمایزکننده برای زبان‌های فارسی، عربی و انگلیسی بود. حال که تنها یک مقدار برای کلیه متون فارسی وجود دارد حذف این ستون اختلالی در پایگاه داده‌ای نمی‌کند.

ستون‌های «آدرس خبر[2]» و «شناسه منبع» با یکدیگر ارتباط نظیر به نظیر دارند؛ بنابراین ستون «آدرس خبر» حذف گردید.

ستون شناسه نوع[3] در جدول خبر دارای شش مقدار متمایز بوده و تنها درباره دو مقدار آن (۰ و ۱) اطلاع داریم. با توجه به کم بودن تعداد رکوردهای باقیمانده، دو راه‌کار داریم: الف) حذف رکوردهای باقیمانده ب) در نظر نگرفتن این ستون در تحلیل داده‌ها.

همچنین در جدول متن تمام رکوردهایی که مقدار ستون «شناسه فایل[4]» در آن‌ها مخالف صفر است حذف شدند؛ زیرا این رکوردها صرفاً برچسب تصاویر خبرها هستند و حاوی محتوای معناداری برای تحلیل نخواهند بود.

```
DELETE
FROM News_Text
WHERE FileId != '0'
```

الگوریتم (۵): حذف خبری که شناسه فیلد آن صفر است

برای برقراری ارتباط بین جداول خبر و متن، از فیلد مشترک آن‌ها (شناسه خبر) در جدول پیوند استفاده شد. در جدول پیوند تمام رکوردهایی ظاهر می‌شوند که شناسه خبر مشترک بین جداول خبر و متن را داشته باشند؛ بنابراین «تعداد رکوردهای پیوند دو جدول» با «تعداد رکوردهای پیوند دو جدول بعد از حذف» با یکدیگر برابر هستند.

[1] Join

[2] URL
[3] Type ID
[4] File ID

```
SELECT *
FROM dbo.News_Text t1
INNER JOIN dbo.News_News t2
ON t1.NewsId = t2.NewsId
```

الگوریتم (۶): واکشی متن خبر با شناسه مشترک

تمامی تغییرات صورت گرفته جهت شناسایی و پیش‌پردازش داده‌های موجود در جدول زیر به نمایش درآمده است.

جدول (۱): نتایج حاصله از پردازش صورت گرفته اخبار

تعداد	توضیح
۱۶۳۴۸۶	تعداد رکوردهای جدول خبر (سطر یکتا: شناسه خبر)
۴۴۷۴۰۷	تعداد رکوردهای جدول متن (سطر یکتا: شناسه متن)
۴۴۷۳۹۲	تعداد رکوردهای پیوند دو جدول
۳۳۶۷	تعداد شناسه خبرهایی که در جدول خبر هستند ولی در جدول متن پیدا نشدند
۱۵	تعداد شناسه خبرهایی که در جدول متن هستند ولی در جدول خبر به آن‌ها ارجاعی داده نشده است.
۱۶۰۱۲۵	تعداد رکوردهای جدول متن که شناسه خبر یکتا دارند.
۱۶۰۱۱۹	تعداد رکوردهای جدول خبر بعد از حذف شناسه خبر ناموجود در جدول متن
۴۴۷۳۹۲	تعداد رکوردهای جدول متن بعد از حذف شناسه خبر ناموجود در جدول خبر
۴۴۷۳۹۲	تعداد رکوردهای پیوند دو جدول بعد از حذف
۷۴۶۲	تعداد رکوردهای جدول متن که حاوی متن انگلیسی می‌باشند.
۱۸۱۰۹	تعداد رکوردهای جدول متن که حاوی برچسب تصاویر بودند.
۴۱۸۸۷۳	تعداد رکوردهای نهایی تا پایان مرحله ۹

جدول حاصل از پیوند دو جدول خبر و متن شامل اطلاعات، شناسه خبر، شناسه متن، نوع متن، متن، شناسه گروه، شناسه منبع، وضعیت و شناسه نوع هست. بدین ترتیب تعداد ستون‌های داده از ۲۰ ستون به ۸ ستون کاهش یافت. دلیل عمده این کاهش بعد نداشتن اطلاعات کافی از ستون موردنظر و یا افزونگی داده‌ای بود.

۱۰- نتیجه‌گیری

در این مقاله، روش ترکیبی‌ای جهت تحلیل داده‌های خبری خبرگزاری‌های فارسی ارائه گردیده است. روش پیشنهادی از استاندارد CRISP تبعیت کرده و در طی انجام مراحل آن از الگوریتم‌های شرح داده‌شده کیفی و کمی داده‌کاوی و الگوریتم Kmeans استفاده می‌کند.

در طی انجام مراحل این تحقیق ابتدا مفاهیم موردنیاز که در انجام پروژه کاربرد دارند به‌صورت کامل شرح داده‌شده‌اند و در مرحله بعد به بررسی سامانه موجود، جهت پیاده‌سازی تحلیل پرداخته می‌شود. فاز بعد به شناخت داده و آماده‌سازی در جهت استفاده در مرحله پیاده‌سازی پرداخته می‌شود. درنهایت در مرحله پیاده‌سازی سعی گردیده است که با کدهای دستوری ساده و قابل‌فهم بر روی پایگاه داده‌ها و روش پیشنهادی پیاده‌سازی این تحقیق که با استفاده از نرم‌افزار RapidMine انجام گردیده به‌صورت کامل شرح داده شود.

استفاده از این روش ترکیبی تحلیل، علاوه بر فراهم کردن تحلیلی مناسب بر روی خبرها و پیدا کردن روابط پنهان‌بین اخبار می‌تواند مزایای فرعی متعددی همچون پایه‌گذاری روشی بومی جهت تحلیل اخبار، روشی سازگار با زبان فارسی، پایه‌گذاری تحلیل‌های معنایی بیشتر بر روی واژگان رایج در اخبار فارسی و درنهایت کمک به تصمیم گیران سازمان‌های نظامی و غیرنظامی جهت انجام تصمیم‌گیری کامل و صحیح را داشته باشد.

۱۱- مراجع

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” AI magazine, vol. 17, p. 37, 1996.

[2] J. Han and M. Kamber, “Data Mining,” Southeast Asia Edition: Concepts and Techniques: Morgan kaufmann, 2006.

[3] M. J. Berry and G. S. Linoff, “Data mining,” techniques: for marketing, sales, and customer relationship management: John Wiley & Sons, 2004.

[4] H. C. Koh and C. K. Low, “Going concern prediction using data mining techniques,” Managerial Auditing Journal, vol. 19, pp. 462-476, 2004.

[5] N. Aggarwal, A. Kumar, H. Khatter, and V. Aggarwal, “Analysis the effect of data mining techniques on database,” Advances in Engineering Software, vol. 47, pp. 164-169, 2012.

[6] D. T. Larose, “k- Nearest Neighbor Algorithm,” Discovering Knowledge in Data: An Introduction to Data Mining, pp. 90-106, 2005.

[7] M. S. Deshpande and D. V. Thakare, "Data mining system and applications: A review," International Journal of Distributed and Parallel systems (IJDPS), vol. 1, pp. 32-44, 2010.

[8] Y. M. Chae, S. H. Ho, K. W. Cho, D. H. Lee, and S. H. Ji, "Data mining approach to policy analysis in a health insurance domain," International journal of medical informatics, vol. 62, pp. 103-111, 2001.

[9] R. Alguliev and R. Aliguliyev, "Experimental investigating the F-Measure as similarity measure for automatic text summarization," Applied and Computational Mathematics, vol. 6, no. 2, pp. 278-287, 2007.

[10] M. E. Califf and R. J. Mooney, "Bottom-up relational learning of pattern matching rules for information extraction," Journal of Machine Learning Research, vol. 4, pp. 177–210, 2003.

[11] M. A. Hearst, "Untangling text data mining," In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics, pp. 3-10, June 1999.

[12] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge discovery and data mining: Towards a unifying framework," In Knowledge Discovery and Data Mining, pp. 82–88, 1996.

[13] V. Kumar and M. Joshi, What is datamining?, 2003. http://wwwusers.cs.umn.edu /~mjoshi/hpdmtut/sld004.htm

[14] R. Feldman and I. Dagan, "Kdt – knowledge discovery in texts," In Proc. Of the First Int. Conf. on Knowledge Discovery (KDD), pp. 112–117, 1995.

[15] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," Communications of the ACM, vol. 18(11), pp. 613–620, 1975. (see also TR74-218, Cornell University, NY, USA).

[16] S. E. Robertson, "The probability ranking principle," Journal of Documentation, vol. 33, pp. 294–304, 1977.

[17] C. J. Van Rijsbergen, "A non-classical logic for information retrieval," The Computer Journal, vol. 29(6), pp. 481–485, 1986.

[18] H. Zhuge et al., "An Automatic Semantic Relationships Discovery Approach," The 13th International World Wide Web Conference (WWW2004), New York, USA, May 2004.

[19] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," In C. Nedellec and C. Rouveirol, editors, European Conf. on Machine Learning (ECML), 1998.

[20] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," In 7th Int. Conf. on Information and Knowledge I, 1998.

[21] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," Machine Learning, vol. 39, pp. 103–134, 2000.

[22] F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys, vol. 34, pp. 1–47, 2002.

[23] N. Kanya and S. Geetha, "Information Extraction," A Text Miningap Proach," produced IEEE, 2007.

[24] H. Karanikas, C. Tjortjis, and B. theodoulidis, "An approach to text mining information exteraction," 2001.

[25] M. Rajman, "Text Mining, Knowledge extraction from unstructured texual data," Proc. of Eurostat Conference, Francfort (Deutchland), May 1997.

Journal of "Innovations of Applied Information and Communication Technologies"
Vol. 1, No. 1, 2019 (Serial No. 1)

Provides a Hybrid Method of Analyzing News Databases Using RapidMine Case Study: Persian News Texts

S. Ahmadian *, K. Saberi

Imam Hossein Comprehensive University

Abstract

One of the most widely used approaches in the management of Web-based systems, use the online agency and Data related to them . Online web agency, like the heart of a community and is one of the most critical information resources in a society. The importance of this topic in the news military organizations will be doubled. So having a right way and a unique conceptual approach to analyze and categorize the resources, Can provide numerous benefits and assistance to decision-makers within the organization, especially military organizations in carrying out decisions. In this study, firstly we contribute to the study and understanding of the concepts of data mining methods. Secondly, analysis of different methods of text processing must be done from different perspectives So as to identify the positive strategies and strengths of a variety of methods and The Finally, the analysis of news database, so as to uncover information hidden in the data and how to use it. Keywords: Data mining, text mining, News Database, Mixed-Method.

Keywords: Data Mining, Text Mining, News Databases

* Corresponding author E-mail: s.ahmadian@ihu.ac.ir