استناد: خوشیان، ناهید؛ امیر غایبی (۱۳۹۷). «درآمدی بر استخراج اطلاعات و استخراج مفاهیم در داده‌کاوی و تفاوت این دو فرایند با یکدیگر (روش‌ها، کاربردها و چالش‌ها) با تأکید بر کاربرد داده‌کاوی در سازمان‌های پلیسی و قضایی به‌منظور کشف جرایم»، توسعه سازمانی پلیس، شماره ۶۶، صص ۱۱۱-۱۵۸.

درآمدی بر استخراج اطلاعات و استخراج مفاهیم در داده‌کاوی و تفاوت این دو فرایند با یکدیگر (روش‌ها، کاربردها و چالش‌ها) با تأکید بر کاربرد داده‌کاوی در سازمان‌های پلیسی و قضایی به‌منظور کشف جرایم

تاریخ دریافت مقاله: ۹۹/۰۱/۲۱
تاریخ پذیرش مقاله: ۹۹/۰۴/۱۳

ناهید خوشیان[۱]، امیر غایبی[۲]

چکیده:

معرفی دو فرایند مهم داده‌کاوی و تحلیل متن- «استخراج اطلاعات و مفهوم»- و بیان تفاوت‌های این دو فرایند با یکدیگر با تأکید بر کاربرد داده‌کاوی در ناجا به منظور کشف جرایم است. در این پژوهش، با استفاده از روش کتابخانه‌ای و بررسی متون و اسناد به گردآوری اطلاعات در مورد استخراج اطلاعات و استخراج مفهوم، روش‌ها، کاربردها و چالش‌های این دو فرایند پرداخته شد. در همین راستا، انواع استخراج اطلاعات و برخی از سامانه‌های استخراج اطلاعات و نیز معماری نظام استخراج اطلاعات و معماری پیشنهادی برای استخراج مفهوم ارائه شده و تفاوت این دو فرایند و روش‌ها، چالش‌ها و کاربردهای گوناگون آن شرح داده شد. نتایج پژوهش نشان می‌دهد که استخراج اطلاعات و استخراج مفهوم دو مبحث مهم در داده‌کاوی و تحلیل متن هستند. استخراج اطلاعات، بازیابی براساس کلیدواژه‌هاست، درحالی‌که استخراج مفاهیم به کلیدواژه‌ها ارتباطی ندارد و با استخراج مفهوم و معنایی که کاربر از متن استنباط می‌کند، ارتباط دارد که این معنا ممکن است لزوماً در کلیدواژه‌ها نباشد. می‌توان گفت استخراج اطلاعات و استخراج مفاهیم، فرایندی چرخه‌ای است که این چرخه مدام غنی و پویا می‌شود. همچنین برمبنای نتایج پژوهش، داده‌کاوی کاربردهای گوناگونی در صنایع مختلف تولیدی، خدماتی و به‌ویژه کاربردهای گوناگونی برای ناجا دارد.

کلیدواژه‌ها:

استخراج اطلاعات، استخراج مفهوم، داده‌کاوی، تحلیل متن، کشف جرایم، سازمان پلیسی و قضایی.

۱. دانشجوی دکترا، دانشگاه الزهرا nkhooshian@gmail.com
۲. دانشیار دانشگاه الزهرا

مقدمه

فناوری اطلاعات موجب ایجاد نظام‌هایی می‌شود که برای انجام وظایفی که عملکردشان به سطح انسانی نزدیک است، کاربرد دارد. با وجود این، چنین نظام‌هایی بدون دارا بودن جامعیت و مانعیت نمی‌توانند به‌خوبی عمل کنند. در چنین مواردی وجود تکنیک‌هایی برای کاهش خطای سیستم در بازیابی و دقت بیشتر توسط کارشناسان ضروری است. وجود نظام‌های استخراج اطلاعات به‌ویژه در موارد گستردگی میزان اطلاعات برای کاربرانی که به این اطلاعات نیاز دارند، بسیار حائز اهمیت است. وجود یک نظام استخراج اطلاعات حتی اگر نتواند نتایج مرتبط با نیاز کاربر را بازیابی کند، نیز بسیار مهم است. استخراج اطلاعات و استخراج مفهوم که در ادامه به آنها پرداخته می‌شود هر دو زیرمجموعه داده‌کاوی و تحلیل متن هستند. داده‌کاوی، فرایندی خودکار برای استخراج الگو و بازنمودن دانش است که این دانش به صورت ضمنی در پایگاه داده‌های عظیم، انباره داده‌ها و دیگر مخازن بزرگ اطلاعات ذخیره شده است. داده‌کاوی، مهم‌ترین مرحله فرایند کشف دانش و هدف آن، کشف دانش است. به‌عبارت دیگر، داده‌کاوی فرایند کشف دانش است و شامل مجموعه‌ای از شیوه‌های ریاضی، رایانه و الگوریتم‌های ویژه است که براساس تحلیل داده‌های موجود در پایگاه داده، برای یافتن راه‌حلی براساس الگوهای کشف‌شده در داده‌ها برمبنای محدودیت‌های مؤثر محاسباتی پذیرفته شده است. درواقع داده‌کاوی، یکی از مهم‌ترین روش‌هایی است که به‌وسیله آن، الگوهای مفید در درون داده‌ها با حداقل دخالت کاربران شناخته می‌شود و اطلاعاتی در اختیار کاربران و تحلیلگران قرار می‌دهد تا براساس آنها، تصمیمات مهم و حیاتی در سازمان‌ها اتخاذ شود. داده‌کاوی را می‌توان فرایندی چهارمرحله‌ای تعریف کرد:

۱- جمع‌آوری مجموعه‌ای از داده‌ها برای تحلیل،

۲- ارائه این داده‌ها به نرم‌افزار داده‌کاوی،

۳- تفسیر نتایج

۴- به‌کارگیری نتایج برای مسئله یا موقعیت‌های جدید.

تحلیل متن یا متن‌کاوی نیز، مجموعه به‌هم‌پیوسته از فناوری‌هایی است که برای پردازش و تحلیل انواع داده‌های نیم‌ساخت‌یافته و غیرساخت‌یافته به‌کار می‌رود و سعی دارد که حروف و واژگان را به عدد تبدیل کند. درحقیقت متن‌کاوی، به تحلیل هوشمند متن، داده‌کاوی متنی یا کشف دانش در متن معروف است و به فرایند استخراج دانش و اطلاعات موردعلاقه و مهم از مجموعه متنی غیرساختاریافته اشاره دارد. تفاوت داده‌کاوی با متن‌کاوی در این است که داده‌کاوی فقط با تحلیل داده‌های عددی سروکار دارد

درحالی‌که متن‌کاوی هم با داده‌های عددی و هم با داده‌های متنی ارتباط دارد (حریری، ۱۳۹۰؛ کاظمی و حسین‌پور، ۱۳۸۸ ؛ لک و رضایی‌نور، ۱۳۹۲).

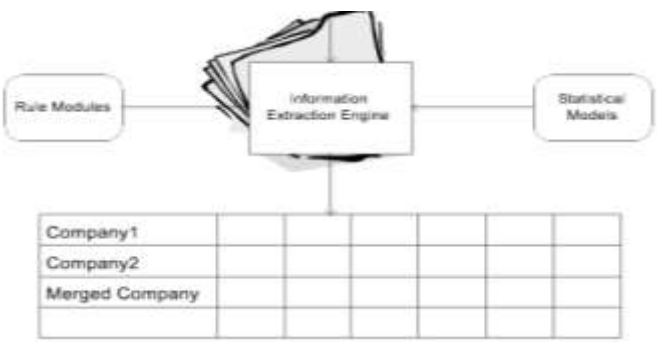
ضرورت وجود نظام‌های استخراج اطلاعات

نظام‌های استخراج اطلاعات به دلیل شرایط زیر ضروری هستند. اطلاعات استخراج‌شده به صراحت مشخص شده است و هیچ‌گونه استنتاج بیشتری موردنیاز نیست. تعداد کمی از قالب‌ها[3] برای خلاصه کردن بخش‌های مربوط به سند کفایت می‌کنند. اطلاعات موردنیاز، در متن بیان می‌شود. به‌عنوان اولین گام، در تَگ‌کردن و برچسب‌دهی به اسناد و مدارک در نظام‌های تحلیل متن، هر سند و مدرک به‌منظور مشخص کردن موجودیت‌ها و روابط معنادار بین آنها، باید پردازش شود. اصطلاح روابط یا ارتباطات در اینجا شامل حقایق یا رویدادهایی با موجودیت‌های مسلم و قطعی است. برای نمونه، یک رویداد ممکن است ورود شرکت‌ها در یک خطر مشترک به‌منظور توسعه یک داروی جدید باشد. یک نمونه از حقیقت، دانشی است در مورد یک ژن که یک بیماری خاص را ایجاد می‌کند. حقایق[4] ایستا هستند و معمولاً تغییر نمی‌کنند. رویدادها بیشتر پویا بوده و دارای مهر تأیید خاص مربوط به خود هستند. استخراج اطلاعات، اطلاعات مختصر و دقیق و همچنین مفاهیم و روابط معنادار که به‌طور مستقیم به دامنه و حوزه[5] مورد بررسی مربوط می‌شوند، فراهم می‌کند. روش‌های استخراج اطلاعات، امکان استخراج اطلاعات واقعی در متن و نه مجموعه محدودی از برچسب‌های مرتبط با اسناد را فراهم می‌کنند. استخراج اطلاعات، موجب ایجاد تعداد محدودی از موجودیت‌ها و روابط است که در آن استخراج متن بدون محدودیت انجام می‌شود. در دهه‌های اخیر اطلاعات متنی در اینترنت رشد سریعی داشته و بخش چشمگیری از این اطلاعات (اخبار برخط، مقالات علمی و کتب و ....) به صورت غیرساخت‌یافته و ناهمگن بوده و اطلاعات غیرساخت‌یافته خواندنی، سازماندهی و تحلیل توسط ماشین‌ها نیستند. برای اینکه بتوان از بین این حجم عظیم اطلاعات، انسان را در فهم و یافتن اطلاعات موردنیاز یاری کرد، باید بتوان متن غیرساخت‌یافته را به اطلاعات ساخت‌یافته تبدیل کرد. درنتیجه وجود فناوری استخراج اطلاعات الزامی است. نظام‌های استخراج اطلاعات با تبدیل اطلاعات به صورت ساخت‌یافته، فهم آن را برای ماشین آسان و به انسان در درک بهتر این اطلاعات کمک می‌کنند.

3. Templates
4. Facts
5. Domain

مثالی از استخراج اطلاعات

برای مثال در جمله زیر John is a graduate student at the University of Pennsylvania سامانه استخراج اطلاعات John، را به‌عنوان Person و University of Pennsylvania را به‌عنوان University برمی‌گرداند. Student-of نیز رابطه بین John و University of Pennsylvania می‌باشد. Person و Student-of نیز، برچسب‌هایی است که از پیش توسط طراح سامانه تعیین شده‌اند. دو وظیفه اصلی استخراج اطلاعات، شامل استخراج موجودیت‌ها و استخراج روابط بین موجودیت‌هاست. تکنیک‌های پیش‌پردازش، که شامل استخراج اطلاعات می‌باشند، قصدشان بر این است که مدل‌های غنی‌تر و انعطاف‌پذیرتری را برای اسناد و مدارک در نظام‌های تحلیل متن ایجاد کنند (فلدمن و سانجر[6]، ۲۰۰۶). برای استخراج اطلاعات، ویژگی‌ها به‌طور مستقیم از متن استخراج می‌شوند. ساده‌ترین نوع استخراج اطلاعات، استخراج اصطلاح[7] است. در شکل زیر تصویری از فرایند استخراج اطلاعات ارائه شده است. در مرکز این روند، موتور استخراج اطلاعات وجود دارد که مجموعه‌ای از اسناد را به‌عنوان ورودی دریافت می‌کند و موتور با استفاده از یک مدل آماری، یک ماژول قانون یا ترکیبی از هر دو کار می‌کند. خروجی موتور مجموعه‌ای از چارچوب‌های حاشیه‌ای استخراج‌شده از اسناد است. عملیات اصلی نظام‌های استخراج اطلاعات از دو مرحله ساخت پایگاه دانش موردنیاز برای استخراج اطلاعات و استفاده از پایگاه دانش برای استخراج اطلاعات از متون ورودی تشکیل شده است (تیمورپور، علی‌زاده و غضنفری، ۱۳۹۵).



شکل ۱. نمایی از نظام استخراج اطلاعات

6. Feldman & Sanger
7. Term extraction

تعریف استخراج اطلاعات

استخراج اطلاعات[8] به عملیات استخراج خودکار اطلاعات ساختاریافته از اسناد و مدارک خواندنی و بدون ساختار یا نیمه‌ساختاریافته اطلاق می‌شود. بخش زیادی از فعالیت‌های استخراج اطلاعات مربوط به پردازش متون توسط روش پردازش زبان‌های طبیعی است. امروزه فعالیت‌های مربوط به پردازش اسناد چندرسانه‌ای مانند حاشیه‌نویسی خودکار، استخراج متون و مطالب از تصاویر، فایل‌های صوتی و کلیپ‌های ویدیویی، از تکنیک‌های استخراج اطلاعات به‌شمار می‌آیند. استخراج خودکار اطلاعات از منابع غیرساختاریافته، راه‌های جدیدی را برای جستجو، سازماندهی و تحلیل داده‌ها با بهره‌گیری از علم معناشناسی از پایگاه داده‌های ساختاریافته و داده‌های غیرساختاریافته گشوده است. (کرمی و ملک‌جعفریان، ۱۳۹۳) پیدایش رشته استخراج اطلاعات در پیکره جامعه پردازش زبان طبیعی اتفاق افتاده است. جایی‌که انگیزه اولیه پیرامون تشخیص موجودیت‌های اسامی نظیر اسامی افراد و سازمان‌ها، از مقالات خبری نشئت می‌گیرد. زمانی‌که جامعه با دسترسی پیوسته به داده‌های ساختاریافته و غیرساختاریافته به سمت داده‌گرایی پیش رفت، کاربردهای جدیدی از استخراج ساختار پدید آمد. پیشینه استخراج اطلاعات به اواسط دهه ۱۹۸۰ بازمی‌گردد، که نظام معاملاتی جایگزین با نام تجاری جاسپر[9] برای رویترز ساخته شد. این نظام با هدف ارائه اخبار مالی در زمان واقعی به معامله‌گران مالی طراحی شده بود. (حاصلی، حسینی‌بهشتی و پاک‌نهاد، ۱۳۹۵).

انواع روابط بین موجودیت‌ها در استخراج اطلاعات

چنانچه پیشتر آمد، دو وظیفه اصلی استخراج اطلاعات، «استخراج موجودیت‌ها و استخراج روابط بین موجودیت‌هاست». به بیان دیگر تشخیص و طبقه‌بندی روابط ازپیش‌تعریف‌شده بین موجودیت‌های مشخص شده در متن. انواع روابط بین موجودیت‌ها عبارت‌اند از:

❖ رابطه بین یک فرد و یک سازمان. مانند:

Steve Jobs works for Apple
Employee of (Steve Jobs, Apple)

❖ رابطه بین یک فرد و یک محل. مانند:

Mr. Smith gave a talk at the conference in New YorkLocated In (Smith, New York)

8. Information extraction
9. JASPER

❖ رابطه بین دو شرکت. مانند:

Listed broadcaster TVN said its parent company, ITI Holdings, is considering various options for the potential sale. Subsidiary Of (TVN, ITI Holding)
(رشادت، حورعلی، ۱۳۹۳).

انواع استخراج اطلاعات

استخراج آزاد اطلاعات

استخراج هدفمند اطلاعات به مشخص کردن نوع اطلاعات موردنظر نیاز دارد، اما در استخراج آزاد اطلاعات با تعدادی نمونه آغازین، مربوط به هر رابطه یا بدون هیچ داده‌ای برای آموزش، اطلاعات را استخراج می‌کنند. استخراج آزاد اطلاعات، یکی از روش‌های استخراج رابطه است و روشی است که برای استخراج نمونه‌های رابطه در متون بزرگ مانند وب مورد استفاده قرار می‌گیرد و برخلاف روش‌های پیشین استخراج اطلاعات، استخراج همه روابط دلخواه از جملات موجود در متن را فراهم می‌کند. به‌عبارت‌دیگر، در برخی موارد هدف کشف تمام حقایق مفید و برجسته موجود در متون بزرگ و متنوع از جمله وب است که به استخراج اطلاعات آزاد نیاز دارد و این روش اولین بار توسط بانکو معرفی شد. در دهه‌های اخیر شاهد انتشار سریع اطلاعات متنی در منابع‌شمار روی اینترنت هستیم. بیشتر آنها اطلاعات متنی غیر-ساخت‌یافته‌اند و جستجو در آنها مشکل است. این نیاز به رویکردهایی برای کشف دانش با ارزش آنها به صورت ساخت‌یافته را نشان می‌دهد که به ظهور فناوری استخراج اطلاعات منجر می‌شود. بنابراین استخراج اطلاعات، شامل شناسایی مفاهیم ازپیش‌تعریف‌شده و نادیده گرفتن اطلاعات بی‌ربط است که با تولید مجموعه عظیمی از اطلاعات متنی، استخراج اطلاعات یک میدان مهم و درحال‌رشد است. به بیان دیگر، استخراج آزاد اطلاعات روشی است که برای کشف روابط از متون بزرگ مانند وب استفاده می‌شود. درواقع در این روش به نوع رابطه خاص اشاره نمی‌شود و برخلاف روش‌های پیشین به مجموعه کوچک از روابط در متن محدود نمی‌شود بلکه همه انواع وابستگی‌های دودویی موجود در متن را استخراج می‌کند و در این راستا از روش‌های بدون ناظر بهره می‌برد (اسماعیلی، ۱۳۹۶).

سامانه‌های استخراج آزاد اطلاعات

در مسیر استخراج اطلاعات مجموعه‌ای از سامانه‌ها برای کمک به استخراج به‌ویژه در زمینه استخراج اطلاعات آزاد معرفی شده‌اند که به شرح زیر است.

سامانه تکست‌رانر[10]: اولین سامانه‌ای بود که با معرفی الگوواره استخراج اطلاعات آزاد عرضه شد. این سامانه با اجرای تعدادی قانون روی داده‌ها، برای خود تعدادی نمونه صحیح ایجاد کرده و سپس آنها را یاد می‌گیرد. این روش را «خودناظر» نام‌گذاری کرده‌اند. سپس از این ابزار برای استخراج رابطه از داده‌ها استفاده می‌شود. سامانه‌ای است که مجموعه‌ای بزرگ از سطرها را بدون نیاز به ورودی انسان استخراج می‌کند.

سامانه ریورب[11]: این سامانه یکی دیگر از سامانه‌های استخراج آزاد اطلاعات است که فعل‌های موجود در متن را می‌یابد و سپس رابطه متناسب با هر فعل را استخراج می‌کند. تجزیه‌کننده نحوی را برای برچسب‌گذاری جملات استفاده می‌کند و محدودیت‌های واژگانی و نحوی را برای شناسایی واقعیات دودویی به‌کار می‌برد.

سامانه اُلی[12]: سامانه اُلی برای استخراج آزاد اطلاعات، ابتدا مجموعه سطرهایی از سامانه ریورب را با خودراه‌انداز مجموعه آموزشی بزرگ به‌کار می‌برد و قالب‌های الگوی باز را روی این مجموعه آموزشی یاد می‌دهد که این قالب‌های الگو در زمان استخراج به‌کار می‌روند. اُلی بهترین شکل عبارت رابطه را برمبنای قالب‌هایی روی عبارت رابطه ریورب تولید می‌کند.

سامانه وُ[13]: این سامانه با استفاده از داده‌های ساخت‌یافته‌ای که در صفحات ویکی‌پدیا وجود دارد، داده‌های موردنیاز را برای آموزش ایجاد می‌کند. این سامانه از خودراه‌انداز مبتنی‌بر ویکی‌پدیا استفاده می‌کند و دسترسی به عبارت رابطه ندارد. این سامانه محدودیت‌های معنایی- لغوی برای الگوها قرار نمی‌دهد و برای عبارت‌های رابطه که فعل میانجی‌شده دارد و شامل اسم نیست، طراحی شده است.

سامانه کراکن[14]: سامانه استخراج آزاد اطلاعات کراکن، برای گرفتن حقایق کامل از جملات عرضه شد و می‌تواند حقایق یک‌تایی، دوتایی تا چندتایی را استخراج کند.

سامانه واندرلست[15]: این سامانه با استفاده از گرامر سبک وابسته عمل می‌کند. مسیرهای وابسته را مطابق قواعد دستوری معتبر برای یافتن آرگومان‌های مرتبط با رابطه پیمایش می‌کند.

سامانه اسنوبال[16]: سامانه نیمه‌نظارتی برای استخراج اطلاعات است که با تعدادی داده آموزشی شروع به کار کرده و سعی می‌کند الگوهای مربوط به وقوع‌های متفاوت این نمونه‌ها را بیابد.

10. Textrunner
11. Reverb
12. OLLIE
13. WOE
14. KRAKEN
15. Wanderlust
16. Snow ball

سامانه نوایت‌آل[17]: برخلاف اسنوبال که به داده ابتدایی برای شروع فرایند استخراج اطلاعات نیاز ندارد، نوایت‌آل برای شروع کار خود به تعدادی الگو و شرح داده موردنظر برای استخراج نیاز دارد. این الگوها وابسته به زبان و البته مستقل از رابطه هستند. سامانه با استفاده از الگوها و داده موردنظر تعدادی عبارت تولید می‌کند و با استفاده از موتور جستجو، صفحات وب مربوط به آن را بازیابی می‌کند و درنهایت اطلاعات از این صفحات بازیابی‌شده استخراج می‌شود (حاصلی، حسینی‌بهشتی و پاک‌نهاد، ۱۳۹۵).

کاربردهای استخراج آزاد اطلاعات

برای استخراج آزاد اطلاعات ۵ کاربرد وجود دارد که عبارت‌اند از: استخراج دانش، استنتاج از زبان طبیعی، استخراج خودکار از شبکه معنایی، پاسخ به پرسش‌ها و مدل زبانی رابطه‌ای. دو کاربرد بسیار برجسته و حائزاهمیت عبارت‌اند از:

۱- استخراج دانش: یکی از روش‌های استخراج دانش، استفاده از منابع دانش خارجی نظیر ویکی‌پدیا برای این کار است. به‌عنوان اولین مرحله استخراج دانش از متن، نیاز است تا موجودیت‌های متن را استخراج کرده و نام‌های هم‌معنا که به یک موجودیت مربوط می‌شوند، با هم در یک گروه قرار دهیم. البته هر نام ممکن است در چند گروه قرار گیرد. برای نمونه نام باراک اوباما[18] می‌تواند در گروه‌هایی متفاوت همراه با رئیس‌جمهور و سیاست‌مدار قرار گیرد. این کار بسیار مشابه با مسئله یادگیری هستان-شناسی خواهد بود. روش‌هایی وجود دارد که برای این کار از منابع دانش خارجی نظیر وردنت یا ویکی‌پدیا استفاده می‌کنند که به‌عنوان معروف‌ترین آنها می‌توان از یاگو[19] و ویکی‌تاکسونومی[20] نام برد. یاگو با استفاده از وردنت و همچنین دسته‌های موضوعی ویکی‌پدیا، روشی برای تولید پایگاه دانش به‌طور خودکار ارائه کرده است.

۲- استنتاج از متن زبان طبیعی: در استنتاج از متن زبان طبیعی، مسئله تشخیص این است که آیا یک جستجو که به زبان طبیعی بیان شده است، می‌تواند به‌طور منطقی از متون موجود که آنها نیز به زبان طبیعی بیان شده‌اند استخراج شود. استنتاج یکی از مباحث اصلی در هوش مصنوعی است که در پنج دهه اخیر سپری‌شده از عمر این رشته، پیشرفت‌های عظیمی در توسعه روش‌های خودکار استنتاج منطقی حاصل شده است. اما چالش استنتاج از زبان طبیعی، همچنان چالشی است کاملاً متفاوت از آنچه که در

17. Know-it-all
18. Barack Obama
19. Yago
20. Wikitaxonomy

این زمینه رخ داده است. مطالعات وسیعی در انواع تحقیقات مرتبط شامل پردازش با زبان طبیعی و بازیابی اطلاعات انجام شده است. با توجه به اینکه بسیاری از اطلاعات ازدست‌رفته در شکل متن باز روی صفحات وب دردسترس است، پس برای استخراج رابطه، روش‌های پردازش متن الزامی است. به همین منظور روش‌های کلی استخراج رابطه شامل مهندسی دانش و یادگیری ماشین معرفی می‌شوند که روش‌های یادگیری ماشینی عبارت‌اند از: باناظر، بدون ناظر و نیمه‌ناظر. روش‌های مهندسی دانش نیز شامل استخراج مبتنی‌بر قالب و استخراج مبتنی‌بر الگو می‌شود (ترکیان، ۱۳۹۷).

روش‌های موجود در استخراج آزاد اطلاعات

در استخراج آزاد اطلاعات که جزو روش‌های یادگیری ماشینی است، از سامانه‌های باناظر، نیمه‌ناظر و بی‌ناظر استفاده می‌شود که در ادامه به توضیح آنها پرداخته می‌شود.

روش‌های یادگیری ماشینی

۱- سامانه‌های بی‌ناظر: سامانه‌های بی‌ناظر به داده‌های آموزشی نیازی ندارند و سعی در تولید آنها ندارند. نام روش بی‌ناظر، رابطه محکمی با خوشه‌بندی دارد و می‌توان آن را رکن اصلی سامانه‌هایی دانست که به صورت بی‌ناظر استخراج می‌کنند. این سامانه‌ها با خوشه‌بندی اخبار براساس مدل رخدادها بگ‌آف-وردز[۲۱]، خبرهای مربوط به یک اتفاق با زمان مشخص را در یک دسته قرار می‌دهند. سپس با خوشه‌بندی دسته‌های حاصل، براساس الگوهای نحوی موجودیت‌های اسامی آنها، خوشه‌های وقایع را ایجاد می‌کند. بدین‌ترتیب حادثه‌ای مثل وقوع طوفان با الگوهای سرعت طوفان و میزان تلفات شهر شناخته می‌شود. به بیان دیگر، این روش برای بهبود معایب روش‌های پیشین استخراج رابطه در متون بزرگ معرفی شد. در روش بدون ناظر، هدف از استخراج اطلاعات، القاء ساختار اطلاعات به‌عنوان انواع رابطه است که روش اصلی مورد استفاده در اینجا خوشه‌بندی می‌باشد. به‌عبارتی مواردی هست که ما هیچ نوع رابطه خاصی در ذهن نداریم، اما می‌خواهیم انواع رابطه برجسته از یک مجموعه داده‌شده را کشف کنیم. شین یاما و همکاران در پژوهشی به کشف رابطه محدود اشاره می‌کنند. آنها ابتدا بسیاری از مقالات خبری را از منابع مختلف در وب جمع‌آوری می‌کنند. سپس مقالات خبری در مورد یک رویداد یکسان را براساس شباهت لغوی خوشه‌بندی می‌کنند. در این روش آنها می‌توانند ویژگی‌های موجودیت را براساس وقوع آن در مقالات توسعه دهند. سپس تجزیه نحوی را انجام می‌دهند و موجودیت‌های اسمی را از مقالات مختلف

21. Bag of Words

استخراج می‌کنند. درنهایت موجودیت‌هایی که همزمان در همان مقاله رخ داده‌اند، براساس ویژگی‌هایشان خوشه‌بندی می‌کنند.

۲- سامانه‌های نیمه‌ناظر: سامانه‌های نیمه‌ناظر مشکل تهیه نمونه‌های آغازین را ندارند. ظهور سامانه‌هایی که با استفاده از چند نمونه آغازین، الگوهای وقوع اطلاعات را یاد می‌گیرند و آنها را استخراج می‌کنند، باعث شد که زحمت انسانی لازم برای تعریف قوانین استخراج کم شود. با وجود تفاوت‌های میان اجراهای مختلف این ایده، می‌توان گفت که همه آنها از روش خودراه‌اندازی استفاده می‌کنند تا نمونه‌های اندک ورودی را با اطلاعات استخراج‌شده مطمئن، افزایش داده و بدین ترتیب، مدل یادگیری را ارتقا دهند. شکل دیگر راهبرد نیمه‌ناظر، یادگیری (مشترک) با هم است که برای اطمینان از سازگاری میان اطلاعات حاصل استفاده می‌شود. در این سامانه، روش‌های مختلفی برای استخراج اطلاعات، در یک فرایند خودراه-اندازی استفاده شده و نتایج هر کدام از روش‌ها موجب تقویت و تصحیح عملکرد دیگران می‌شود. در ادامه به بررسی روش‌های نیمه‌ناظر پرداخته خواهد شد.

روش خودراه‌انداز : روش‌های مبتنی‌بر هسته و مبتنی‌بر ویژگی از داده‌های بسیار زیاد آموزشی برای استخراج اطلاعات استفاده می‌کنند. برای حل این مسئله روش یادگیری نیمه‌ناظر معرفی شد که با داده آموزشی کم کار می‌کند. روش مهم در یادگیری نیمه‌ناظر روش خودراه‌انداز است که از مجموعه کوچک نمونه‌های رابطه شروع می‌شود و از الگوهای استخراج استفاده می‌کند. سامانه اسنوبال توسط آگیچتین و گراوانو برای استخراج اطلاعات در روش خودراه‌انداز عرضه شد. ایده این سامانه ساده است و با جفت موجودیت‌های مرتبط با رابطه هدف شروع می‌کند و در متن به جستجوی جفت موجودیت‌های مجاور می‌پردازد. اگر جفت موجودیت‌ها به‌طور همزمان در متن رخ داده باشند (مفهوم همزمانی موجودیت‌ها احتمالاً به معنی الگویی برای رابطه هدف است)، پس جفت موجودیت‌ها به نمونه رابطه اضافه می‌شوند و تا زمانی‌که شرایط دقیقی ایجاد شود، پردازش ادامه دارد؛ به‌طوری‌که بیشتر الگوها و موجودیت‌ها به نتایج پردازش اضافه می‌شوند. یک گام مهم در شیوه خودراه‌انداز ارزیابی کیفیت الگوهای استخراج است. در نتیجه فرایند استخراج شامل الگوهای خراب نمی‌شود.

۳- روش نظارت راه دور: با رشد وب اجتماعی، دانش انسان، بیشتر توسط کاربران در پایگاه‌های اطلاعات ذخیره می‌شود. نمونه کاملاً شناخته‌شده آن ویکی‌پدیاست. در این حالت، ممکن است مجموعه‌ای بزرگ از موجودیت‌ها برای رابطه هدف باشند تا داده آموزشی تولید شود. روش نظارت راه دور ویژگی‌های استخراج‌شده از جملات متفاوت شامل هر جفت موجودیت را برای ایجاد بردار ویژگی غنی استفاده می‌کند.

روش نظارت راه دور یا یادگیری خودنظارتی برای استخراج پایگاه‌های دانش بزرگ برای برچسب زدن خودکار موجودیت‌ها در متن و استخراج ویژگی‌ها و آموزش دادن دسته‌بند به‌کار می‌رود. این روش فقط برای استخراج روابطی که از مرز جملات رد نشده‌اند و جملاتی که حاوی اشاره روشنی از فعل و فاعل رابطه است، استفاده می‌شوند. استخراج نظارت راه دور به‌عنوان برچسب‌گذاری خودکار متن با ویژگی‌ها و منابعی که منابع موجودیت‌ها از یک پایگاه دانش هستند، استفاده می‌شوند. اگر دو موجودیت در یک رابطه باشند، هر جملهٔ شامل این دو موجودیت، ممکن است این رابطه را تبیین کند. پیش از استفاده از برچسب-گذاری خودکار متن برای آموزش دسته‌بند، نمونه‌های حاوی واژگان مبهم کشف و دور انداخته می‌شوند. اولین رویکرد این است که اگر لغات مبهم هستند، آنها از اشیا برای موجودیت هدف تفکیک می‌شوند. درواقع اگر واژگان موضوع در سراسر کلاس مبهم باشد، درنتیجه موضوع درکلاس خاص مبهم است.

۴- سامانه‌های باناظر (خودناظر): سامانه باناظر، داده آموزشی را برای خود تهیه می‌کند. تهیه نمونه برای هر رابطه، اگر چه تعداد این داده‌ها بسیار کم باشد، (روش‌های باناظر با داده آموزشی کم کار می‌کنند)، سامانه را با چالش مناسب بودن داده‌ها مواجه می‌کند. یعنی باید مطمئن باشیم که داده فراهم‌شده، موجب بازیابی بخش خاصی از نتایج نمی‌شود. در مقابل این رویه، سامانه‌های باناظر با ادعای حذف نمونه‌های آغازین تلاش می‌کنند داده‌های موردنیاز خود را تولید کنند و می‌توان آنها را جزو سامانه‌های بی‌ناظر به‌شمار آورد. این روش به جای داده‌های آغازین با تعدادی الگو و نوع شروع می‌شود. قرار دادن هر نوع در الگوها، به ایجاد جستجوی مناسب برای بازیابی نمونه‌های آن نوع می‌شود. یعنی نمونه‌ها از طریق انطباق الگو، با نتایج جستجو استخراج می‌شوند. همان‌طور که گفته شد، روش‌های باناظر با داده آموزشی کم کار می‌کنند و به دو دسته اصلی به شرح زیر تقسیم می‌شوند.

۱- دسته‌بندی براساس ویژگی: این روش معمول استخراج رابطه است که مشکل دسته‌بندی را حل می‌کند. به‌طور خاص، هر جفت موجودیتی که درجمله اتفاق می‌افتد، به‌عنوان نامزد مطرح می‌شود. هدف، تخصیص برچسب کلاس به جفت موجودیت است که برچسب کلاس یک رابطه ازپیش‌تعریف‌شده برای جفت موجودیت نامرتبط است. مهندسی ویژگی گام مهمی در روش دسته‌بندی است.

۲- روش هسته: مهمترین کار در استخراج رابطه، دسته‌بندی براساس هسته است. در یادگیری ماشینی، یک تابع هسته یا کرنل محصول داخلی نمونه‌های مشاهده‌شده را در بعضی زیرلایه‌های فضای برداری تعریف می‌کند. مزیت عمده استفاده از هسته این است که موارد مشاهده‌شده برای محاسبه شدن

لازم نیست به صراحت به فضای برداری محصولات داخلی خود نگاشت شود (حاصلی، حسینی‌بهشتی و پاک‌نهاد، ۱۳۹۵).

روش‌های مهندسی دانش

برای استخراج دانش و اطلاعات از متن نیاز است تا موجودیت‌های متن را استخراج کرده و نام‌های هم‌معنا که به یک موجودیت مربوط می‌شوند، با هم در یک گروه قرار دهیم. البته هر نام ممکن است در چند گروه قرار گیرد. برای نمونه نام حسن روحانی می‌تواند در گروه‌هایی متفاوت همراه با نام رئیس‌جمهور، سیاست‌مدار و استاد دانشگاه قرار گیرد. این کار بسیار مشابه با مسئله یادگیری هستان‌شناسی خواهد بود. روش‌هایی وجود دارند که برای این کار از منابع دانش خارجی نظیر وردنت یا ویکی‌پدیا استفاده می‌کنند که بهعنوان معروفترین آنها می‌توان از ویکی و یاگو نام برد. یاگو با استفاده از وردنت و همچنین دسته‌های موضوعی ویکی‌پدیا روشی را برای تولید پایگاه دانش بهطور خودکار ارائه کرده است. روش‌های مهندسی دانش به دو دسته مبتنی‌بر قالب و مبتنی‌بر الگو تقسیم می‌شود که شرح آن در ادامه می‌آید.

۱- روش مبتنی‌بر قالب : منظور از قالب، نحوه بیان یک واقعه به همان شکلی است که معمولاً بیان می‌شود. در بیان هر واقعه‌ای معمولاً تعدادی نقش معنایی در شکل‌های متنوع و البته محدود ظاهر می‌شوند. برای نمونه در یک خبر مربوط به بمب‌گذاری، از نقش عامل بمب‌گذار، منطقه آسیب‌دیده و ... صحبت می‌شود. روشن است که هر قالب حجم بسیاری از اطلاعات را در خود جای می‌دهد و پیشنهاد اقتباس و استفاده از آن به شکل بی‌ناظر کمی عجیب به‌نظر می‌رسد. استخراج اطلاعات به این نحو نیز تجربه شده است. ویژگی‌های به‌دست‌آمده از متون براساس میزان باهم‌آیی آنها خوشه‌بندی می‌شود تا به خوشه‌هایی که هر کدام در مورد موضوع مشخصی صحبت می‌کنند، دست یابیم. پس از این مرحله امیدواریم که مثلاً یک خوشه مربوط به اخبار بمب‌گذاری باشد و خوشه دیگر مربوط به اخبار آدم‌ربایی (در فضای متون خبری پلیسی صحبت می‌شود). متون هم‌موضوع در یک خوشه قرار می‌گیرند تا برای هر کدام از آنها قالبی با نقش‌های معنایی مشخص، کشف شود. یادآوری می‌شود که فرایند کاملاً بی‌ناظر و مبتنی‌بر دانش است. توصیف از نقش معنایی به شکل آرگومان‌های ممکن آنها در هر ویژگی استخراج می‌شود. مثلاً نقش مکان آسیب‌دیده می‌تواند مفعول تخریب کردن ظاهر شود. پس از رسیدن به قالب‌های وقایع، استخراج اطلاعات آنها بسیار آسان است. هر کدام از نقش‌های معنایی، یکی از اطلاعات موردنظر است که برای متون جدید به‌آسانی می‌توان آنها را استخراج کرد.

۲- روش مبتنی‌بر الگو : متن‌کاوی همچنان یکی از موضوعات چالش‌برانگیز است که برای استخراج دانش مفید از متن استفاده می‌شود تا به کاربران در یافتن الگوهای مفید و آنچه می‌خواهند، کمک کند. ازجمله مزایای استفاده از روش‌های مبتنی‌بر ترم، عملکرد محاسباتی مناسب به‌علاوه نظریه‌های کامل برای وزن‌دهی ترم‌ها می‌توان نام برد. اگرچه روش‌های مبتنی‌بر ترم از مشکلات معنایی و هم‌معنایی رنج می‌برند. منظور از معنایی این است که یک کلمه ممکن است چند مفهوم و معنا داشته باشد و منظور از هم‌معنایی این است که چند کلمه ممکن است یک معنی و مفهوم را داشته باشند. در طول استخراج اطلاعات، این فرضیه وجود دارد که همواره رویکردهای مبتنی‌بر عبارت، عملکرد بهتری از رویکردهای مبتنی‌بر ترم دارند. برای نمونه، عبارت، ممکن است معنای بیشتری را مانند اطلاعات با خود حمل کند. عبارت‌ها کمتر مبهم هستند و قابلیت متمایزکنندگی بیشتری نسبت به ترم‌های فردی دارند. اما این فرضیه اکنون شانس چندانی ندارد به چند دلیل: ۱) عبارت‌ها ویژگی آماری نامرغوبی نسبت به ترم‌ها دارند. ۲) تعدد وقوع در آنها بسیار پایین است. ۳) ممکن است تعداد زیادی عبارت تکراری و اضافه وجود داشته باشد. پس الگوهای ترکیبی در حوزه متن‌کاوی به یک جایگزین امیدوارکننده برای عبارت‌ها تبدیل شد. زیرا الگوهای ترتیبی، ویژگی آماری ترم‌ها را به‌خوبی درنظرمی‌گیرند. روش‌های تطبیق الگو، به‌طور وسیعی در قلمروی استخراج اطلاعات کاربرد دارند و در قلمروی یادگیری هستان‌شناسی نیز به ارث رسیده‌اند. در روش‌های مبتنی بر الگو، ورودی (معمولاً متن)، به دنبال الگو یا کلمه کلیدی ویژه که نشانگر رابطه مفهومی خاص است، جستجو می‌شود. این الگوها انواع مختلفی (اعم از نحوی یا معنایی و عمومی یا خاص) دارند و برای استخراج عناصر مختلف هستان‌شناسی مثل روابط طبقه‌ای یا غیرطبقه‌ای یا اصول بدیهی به‌کار می‌روند.

چالش‌های استخراج آزاد اطلاعات

چالش‌های نظام‌های استخراج آزاد روابط شامل این است که این نظام‌ها نمی‌توانند تمام روابط را استخراج کنند و از طرفی خروجی ناقص و نوفه‌دار دارند و نیز ممکن است استخراج اطلاعی دربرنداشته‌باشند. از دیگر مشکلات این نظام‌ها این است که به دلیل ماهیت مقیاس‌پذیر بودن استخراج آزاد روابط، استفاده از ابزارهای عمیق پردازش زبان طبیعی نظیر تجزیه‌گر نحوی و معنایی که موجب بهبود چشمگیر نتایج و افزایش دقت می‌شود، ممکن نیست. از طرفی استفاده صرف از ابزارهای سطحی پردازش زبان طبیعی نظیر تجزیه‌گر سطحی، اجزای سخن و ... باعث کاهش چشمگیری در معیارهای کارایی استخراج‌گرها می‌شود. همچنین استخراج هدفمند اطلاعات نیازمند صرف توان زیاد انسانی برای مشخص

کردن محدوده و نوع دانش مورد تقاضاست و این یعنی رویه هدفمند، توان مواجهه با حجم عظیمی از اطلاعات را ندارد. اما استخراج آزاد اطلاعات، به‌طور دقیق در پی حل این مشکل است. صحبت از روشی برای استخراج اطلاعات کاملاً متنوع از انبوه متون در وب است. ورودی سامانه، حجم زیادی از متن خام و خروجی آن، روابط درون متن است، با همان شکلی که ظاهر شده‌اند. درواقع اطلاعات موردنظر، با یک بار عبور سریع از متن استخراج می‌شوند. از نظر بنکو و همکاران مواجهه با این مسئله، چند چالش اساسی دارد که عبارت‌اند از:

۱- دخالت نکردن انسان: عبور از استخراج هدفمند اطلاعات، آغاز خوبی برای کاهش زحمت تهیه داده‌های آموزشی سامانه بود. با ظهور سامانه‌هایی که تنها به چند نمونه آغازین برای استخراج نمونه‌های مشابه نیاز داشتند، کاربر باید برای هر رابطه موردنظر، چند نمونه تهیه می‌کرد. وقتی تعداد روابط زیاد باشد، همین مسئله مانعی برای استفاده از سامانه در ابعاد وب می‌شود. افزون‌بر این، کیفیت نمونه‌های خروجی کاملاً وابسته به کیفیت ورودی‌هاست. از این‌رو به دنبال روشی هستیم که هیچ نمونه‌ای برای یادگیری یا حتی شروع آن ورودی نگیرد و تمامی روابط موجود را استخراج کند.

۲- ناهمگونی در پیکره: وقتی از یک الگوی باناظر برای استخراج اطلاعات استفاده می‌کنیم، افزون‌بر اینکه به داده آموزشی نیاز داریم، سامانه آموزش‌دیده باید برای داده‌های مشابهی استفاده شود. برای نمونه اگر از ویژگی‌های لغوی در هنگام یادگیری استفاده کنیم، نمی‌توانیم انتظار داشته باشیم سامانه در متونی با دایره لغات متفاوت، به‌خوبی عمل کند. انتظار ما از سامانه استخراج اطلاعات، از تمامی وب است. یعنی عرصه‌ای که ناهمگونی بسیار زیادی در واژه‌ها و سبک نوشتار آن وجود دارد.

۳- کارایی: سامانه‌هایی ارائه شده‌اند که انسان در آنها، تقریباً هیچ نقشی ایفا نمی‌کند و همچنین تا حدی با ناهمگونی داده‌ها کنار می‌آیند. اما بهای زیادی برای این کار می‌پردازند. برای نمونه سامانه‌ای که برای استخراج اطلاعات، به صورت پیوسته از یک موتور جستجو کمک می‌گیرد و همین یعنی بسیار کند عمل می‌کند. همچنین در روش‌های بی‌ناظر باید، حجم عظیم داده‌ها را خوشه‌بندی کنیم که مانع از مواجهه با انبوه متون می‌شود. در استخراج آزاد به دنبال روشی هستیم که با سرعت زیاد و با عبور یکباره از روی متن، اطلاعات موردنظر را به دست آورد (اسماعیلی، ۱۳۹۶).

استخراج هدفمند اطلاعات

چنان‌چه گفته شد روش‌های کلی استخراج رابطه افزون‌بر یادگیری ماشینی شامل مهندسی دانش نیز می‌باشد. استخراج هدفمند اطلاعات جزو روش‌های مهندسی دانش است. استخراج اطلاعات به‌منظور

تبدیل اطلاعات قابل‌استفاده از منظر ماشین انجام می‌شود. رویه مرسوم در پژوهش‌های دیرین مرتبط با این موضوع، استخراج برای اهداف مشخصی بوده است. یعنی ماشین باید نوع خاصی از اطلاعات که احتمالاً به شکل واضحی در متن بیان شده‌اند، استخراج می‌کرده است. استخراج هدفمند اطلاعات به‌عنوان جایگزین مناسبی برای این نوع نگرش به مسئله، به نظر می‌رسد. برای نمونه استخراج زمان و مکان برگزاری کنفرانس‌ها را می‌توان یک استخراج هدفمند دانست که معمولاً باید روی دامنه مشخصی از متن‌ها (مثل اعلان برگزاری کنفرانس‌ها) انجام گیرد. وقتی استخراج اطلاعات را به صورت یک مسئله باناظر تعریف می‌کنیم، درواقع استخراج هدفمند را برگزیده‌ایم. یعنی فرض می‌کنیم که اطلاعات مشخصی برای استخراج، توسط انسان معین شده‌اند و حالا ماشین باید این داده‌ها را یاد بگیرد تا بتواند از متون جدید نیز اطلاعات موردنظر را استخراج کند. اگر چه الزامی وجود ندارد که مجبور به انتخاب رابطه بیا روابطی محدود باشیم، نوع نزدیک شدن ما به مسئله به این محدودیت منجر می‌شود.

روش‌های استخراج هدفمند اطلاعات

سه روش عمده برای استخراج هدفمند اطلاعات وجود دارد.

۱- استخراج اطلاعات مبتنی‌بر قاعده: استفاده از قاعده‌های مشخصی برای استخراج اطلاعات در دامنه‌های محدود، انتخاب خوبی است. برای نمونه سامانه‌های یاگو[۲۲] و دی‌بی‌پدیا[۲۳] با کمک قواعد تهیه‌شده توسط انسان، انبوه اطلاعات ساخت‌یافته موجود در ویکی‌پدیا را استخراج می‌کنند. اطلاعات ساخت‌یافته ویکی-پدیا در قالب اینفوباکس و همچنین فهرست‌ها و جداول ارائه شده‌اند. قاعده‌های موردنظر فاصله میان واژه و اطلاع را تکمیل می‌کنند، برای نمونه باید مشخص کنیم که رابطه متولد شدن با ردیف تاریخ تولد مشخص می‌شود.

۲- مدل‌های گرافی: استفاده از مدل‌های گرافی (مدل گرافی مدلی احتمالاتی است که در آن استقلال شرطی بین متغیرها توسط یک گراف نشان داده می‌شود. دو نوع مدل گرافی وجود دارد: مدل با گراف جهت‌دار که شبکه بیزی نامیده می‌شود و مدل با گراف ساده که میدان تصادفی مارکفی نامیده می‌شود. مدل‌های گرافی به وفور در نظریه احتمالات و آمار به‌ویژه در آمار بیزی و یادگیری ماشین به‌کار می‌روند) برای حل مسائل باناظر کاملاً مرسوم است. مسئله استخراج اطلاعات را نیز می‌توان در این قالب تعریف کرد. برای نمونه استخراج ویژگی‌های مقاله از میان سربرگ و ارجاع‌ها به شکل یک مسئله پیش‌بینی

22. YAGO
23. DBpedia

ساختار تعریف و حل شده است. پس از تعریف مسئله در این قالب، باید ویژگی‌های مورد بررسی برای قضاوت را مشخص کنیم. برای استخراج اطلاعات مقاله، سه نوع ویژگی مورد استفاده قرار گرفته‌اند. گروهی از ویژگی‌ها مربوط به ظاهر واژه‌ها هستند، مثل شروع شدن واژه با حرف بزرگ یا اینکه واژه شکل مختصر یک عبارت است. این نوع ویژگی در بیشتر مدل‌های گرافی مربوط به زبان طبیعی استفاده می‌شود. بخش دیگری از ویژگی‌ها مربوط به چینش و محل قرارگیری واژه هستند. مثل شروع شدن خط یا اتمام آن با واژه. واضح است که این ویژگی به مسئله خاص استخراج اطلاعات از روی مقاله‌ها مربوط است و در همه مسائل نمی‌توان از آن استفاده کرد. گروه دیگری از ویژگی‌ها هم مربوط به منابع دانش خارجی مثل فرهنگ‌ها هستند مانند حضور واژه در فهرست نویسندگان مقاله. انتخاب ویژگی‌ها ممکن است تا حدی سلیقه‌ای باشد و به آزمون و خطا نیاز داشته باشد. اما درمجموع، مدل گرافی بستر خوبی برای استفاده از ویژگی‌های تعریف شده است.

۳- تابع هسته: در یادگیری ماشینی عادت کرده‌ایم برای قضاوت از توصیف برداری داده‌ها استفاده کنیم. یعنی مثلاً برای تشخیص اینکه واژه موردنظر جزئی از اطلاعات نیازمند استخراج است یا نه، مجبوریم واژه را با چند ویژگی، چنان‌که در بخش پیش گفته شد توصیف کنیم. در این صورت امکان استفاده از طیف گسترده ابزارهای رده بند را داریم. تابع هسته این فرصت را فراهم می‌کند که داده‌ها را به صورت بردار ویژگی‌ها توصیف نکنیم و فقط نحوه محاسبه شباهت میان آنها را بیان کنیم. همین ویژگی است که استفاده از تابع هسته را در پردازش زبان طبیعی امیدبخش می‌کند. تعریف هسته‌های مربوط به زبان طبیعی برای مقایسه درخت‌های تحلیل با پیدا کردن زیردرخت‌های مشترک، از نمونه‌های موفق استفاده این روش است. توابع هسته برای استخراج اطلاعات نیز تعریف شده و مورد استفاده قرار گرفته‌اند. برای نمونه استفاده از تجزیه کم‌عمق جمله برای تشخیص رابطه اشخاص و نهادها و همچنین مکان سازمان‌ها بررسی شده است. مزیت اصلی استفاده از تابع هسته در این مورد نیز همان عدم لزوم تبدیل تجزیه کم‌عمق به بردار ویژگی‌هاست.

کاربردهای استخراج هدفمند اطلاعات

ساراواگی[۲۴] کاربردهای استخراج هدفمند اطلاعات را در ۴ گروه کاربردهای تجاری، شخصی، علمی و مبتنی‌بر وب دسته‌بندی کرده است.

24. Sara Vagi

۱- کاربردهای تجاری: شامل پیگیری اخبار، مراقبت از مشتری، تمایز داده‌ها و تبلیغات طبقه‌بندی شده است.

۱-۱- پیگیری اخبار: یکی از ابتدایی‌ترین کاربردهای استخراج اطلاعات است که به صورت خودکار انواع رخدادهای خاص را از منابع خبری ردیابی می‌کند. این رخدادها شامل طیف وسیعی مثل اسامی افراد و شرکت‌ها، روابط ساختاریافته بین موجودیت‌ها (مثل همدان بخشی از ایران)، پیگیری بروز بیماری‌ها و حوادث تارش‌گری (تروریستی) از منابع خبری است.

۱-۲- مراقبت از مشتری: شرکت‌های مشتری‌مدار، بسیاری از اشکال داده‌های ساختارنیافته از تعامل کاربران را گردآوری می‌کنند. برای مدیریت مؤثر، این داده‌ها باید با پایگاه داده‌های ساختاریافته، شرکت و هستی‌شناسی‌های کسب‌وکار ادغام شود. این امر باعث می‌شود بسیاری از مسائل استخراج اطلاعات، نمایان شود. مسائلی همچون شناسایی اسامی محصولات و ویژگی‌های آنها از پیام‌نگار (پست الکترونیک) مشتریان، ارتباط بین پیام‌های مشتریان با یک تراکنش خاص در یک پایگاه داده فروش، استخراج اسامی و نشانی‌های تجاری از فاکتورهای فروش، استخراج سوابق تعمیر از کاربرگه‌های درخواست بیمه و غیره.

۱-۳- تمایز داده‌ها: یکی از مراحل اساسی در تمام فرایند تمیز انبار داده، تبدیل نشانی‌هاست که به صورت حروف در شکل‌های ساختاریافته خود همچون اسامی جاده‌ها، شهرها و ایالت‌ها ذخیره شده‌اند. سازمان‌های مشتری‌مدار بزرگ مانند بانک‌ها، شرکت‌های مخابراتی و دانشگاه‌ها، میلیون‌ها نشانی را ذخیره می‌کنند که در شکل اصلی، این نشانی‌ها، ساختار روشنی ندارند. معمولاً برای یک شخص نشانی‌های مختلفی در پایگاه داده‌های مختلف وجود دارد. در ساختار انباره داده لازم است این آدرس‌ها در یک شکل استاندارد کانونی قرار داده شوند. جایی که گزینه‌های مختلف شناسایی و موارد تکراری حذف شوند. یک شکننده رکورد نشانی در ابعاد ساختاریافته آن، نه‌تنها باعث ایجاد جستجوی بهتر می‌شود، بلکه یک راه قوی‌تر جلوگیری از تکرار و جادهی را فراهم می‌کند.

۱-۴- تبلیغات طبقه‌بندی‌شده: تبلیغات طبقه‌بندی‌شده و دیگر فهرست‌ها نظیر فهرست‌های رستوران-ها، حوزه دیگری با ساختار مشخص هستند که در صورت نمایش می‌توانند برای جستجو بسیار باارزش باشند. بسیاری از پژوهشگران اهداف خاصی نظیر داده‌های رکوردمحور در پژوهش‌های استخراجی خود دارند.

۲- مدیریت اطلاعات شخصی: نظام‌های مدیریت اطلاعات شخصی در جستجوی سازماندهی داده‌های شخصی همچون مدارک، پیام‌نگارها، پروژه‌ها و افراد در یک قالب ساختاریافته با پیوند درونی است. موفقیت این نظام‌ها بستگی به توانایی آنها در استخراج خودکار اطلاعات ساختاریافته از منابع ساختارنیافته مبتنی‌بر

پرونجا (فایل) دارد. بنابراین برای نمونه تونایی استخراج خودکار اطلاعات از پرونجاهای پرده‌نگار، پدیدآور یک گفتگو و پیوند شخص با ارائه‌دهنده گفتگوی اعلان‌شده در پیام‌نگار است.

۳- کاربردهای علمی: پیشرفت اخیر رشته انفورماتیک زیستی، دامنه استخراج از موجودیت‌های اسامی را به اشیاء زیستی همچون پروتئین‌ها و ژن‌ها گسترش داده است. مشکل اصلی استخراج اطلاعات از پایگاه مقالات همچون پاب مد، اسامی پروتئین‌ها و تعاملات بین آنهاست. ازآنجایی‌که قالب موجودیت‌هایی مثل اسامی ژن‌ها و پروتئین‌ها خیلی متفاوت از اسامی موجودیت‌های کلاسیک مثل افراد و سازمان‌هاست، این وظیفه به گسترش فنون مورد استفاده برای استخراج اطلاعات کمک کرده است.

۴- کاربردهای وب‌محور: بسیاری از پایگاه‌های استنادی در وب از طریق مراحل استخراج ساختار استادانه از منابعی همچون وبگاه‌های کنفرانسی تا صفحات خانگی شخصی ایجاد شده‌اند. که از این میان سیر، گوگل اسکالر و کورا محبوب‌ترین آنها هستند.

۱-۴- پایگاه داده‌های نظریه‌ای: وبگاه‌های بی‌شماری وجود دارند که نظرات مورد نظارت قرارنگرفته‌ای در مورد دامنه وسیعی از موضوعات شامل تولیدات علمی، کتاب‌ها، فیلم‌ها، افراد و موسیقی ذخیره می‌کنند. ارزش این مطالب در صورت سازماندهی با دامنه‌های ساختاریافته می‌تواند به میزان زیادی افزایش یابد.

۲-۴- وبگاه‌های جوامع: نمونه دیگر از ایجاد پایگاه داده‌های ساختاریافته از مدارک وب، وبگاه‌های جوامعی همچون دی‌بی‌لایف[۲۵] و رگساست که به ارائه اطلاعات در مورد پژوهشگران، کنفرانس‌ها، پروژه‌ها و رخدادهای مرتبط با یک جامعه خاص پرداخته و آن را پیگیری می‌کند. ایجاد این گونه پایگاه داده‌های ساختاریافته، نیازمند مراحل استخراج بسیاری است: جایابی صحبت‌های ارائه‌شده از صفحات سازمان، استخراج اسامی سخنرانان، استخراج پیشینه‌های ساختاریافته در مورد یک کنفرانس و جزء آن.

۳-۴- خرید مقایسه‌ای: خرید مقایسه‌ای برای ایجاد وبگاه‌های خرید مقایسه‌ای است که به صورت خودکار، وبگاه‌های تجاری را برای یافتن محصولات و قیمت آنها پیمایش می‌کنند. از زمان توسعه فناوری‌های وب، بسیاری از وبگاه‌های تجاری شروع به پنهان کردن قالب‌ها و زبان‌های برنامه-نویسی کردند. درنتیجه کانون تمرکز به پیمایش و استخراج اطلاعات از وبگاه‌های مبتنی‌بر شکل انتقال پیدا کرد. برای نمونه، گنجاندن تبلیغ در صفحات وب. صفحاتی را در نظر بگیریم که می-خواهند محصولی را به صورت پشت‌سرهم و به طریقی که اشاره آنها به محصول و بیان یک نظر

25. DBLIFE

مثبت در مورد آن باشد، تبلیغ کنند. هر کدام از وظایف جزئی استخراج ویژگی‌های محصول و نوع نظر اظهارشده در مورد محصول، نمونه‌هایی از وظایف استخراج اطلاعات هستند. (زاهدی احمدسرایی و مهردوست، ۱۳۹۳)

چالش‌های استخراج هدفمند اطلاعات

چالش بزرگ استخراج اطلاعات، جستجوهای وب ساختاریافته، یعنی اجازه جستجوهای ساختاریافته شامل موجودیت‌ها و روابط آنها در وب، است. جستجوی کلیدواژه‌ای برای ارائه اطلاعات در مورد موجودیت‌هایی که معمولاً اسم یا عبارات اسمی هستند، مناسب بوده اما در مورد جستجوهای کشف روابط بین موجودیت‌ها، ناتوان هستند. جدول زیر به تفکیک روش‌های استخراج اطلاعات و دسته‌بندی هر یک از این روش‌ها و سامانه‌های استخراج اطلاعات را نشان می‌دهد (تیمورپور، علی‌زاده و غضنفری، ۱۳۹۵).

جدول ۱. مقایسه روش‌های استخراج اطلاعات

انواع استخراج اطلاعات	نوع دسته‌بندی	روش استفاده شده	رویکردها
مهندسی دانش	مبتنی‌بر قالب	خوشه بندی متون با موضوع یکسان و استخراج قالب مشخص	استخراج هدفمند اطلاعات
	مبتنی‌بر الگو	استخراج الگو یا کلمه کلیدی خاص از متن، استخراج عناصر مختلف هستی‌شناسی	
یادگیری ماشینی	بدون ناظر	کشف رابطه و القای الگو	استخراج آزاد اطلاعات
		دسته‌بندی براساس ویژگی	
	باناظر	دسته‌بندی براساس هسته	
	نیمه‌ناظر	خودراه‌انداز	
		نظارت راه دور	

چهار نوع از عناصر که می‌توانند از متن استخراج شوند عبارت‌اند از:

موجودیت‌ها[۲۶]: که سازه‌های اصلی هستند و می‌توانند از اسناد و مدارک متن استخراج شوند، شامل افراد، شرکت‌ها، داروها و ژن‌هاست.

26. Entity

ویژگی‌ها[۲۷]: ویژگی‌ها، موجودیت‌های استخراج‌شده هستند. چند نمونه از ویژگی‌ها عبارت‌اند از عنوان یک فرد، سن فرد و نوع سازمان.

حقایق[۲۸]: ارتباطاتی هستند که بین موجودیت‌ها وجود دارند. برای نمونه، رابطه کاری بین شخص و شرکت یا فرایند فوسفوریلاسیون بین دو پروتئین.

رویداد[۲۹]: یک فعالیت یا رخداد است که در آن موجودیت‌ها عبارت‌اند از: یک اقدام تروریستی، ادغام دو شرکت، تولد و ... .

نکته دیگر در استخراج اطلاعات، این است که چگونه موجودیت‌هایی را که در یک مدرک هستند مورد بررسی قرار دهیم. یک گزینه ممکن، استخراج همه آنهاست که با این دید هر گونه حذف کردن یا نادیده گرفتن اطلاعات به جامعیت پایین‌تر منجر می‌شود. یک گزینه دیگر این است که هر موجودیت تنها یکبار استخراج شود، به‌عبارتی شناسایی تنها یک رخداد از هر موجودیت کفایت می‌کند. در بسیاری از موارد گزینه دوم یعنی اینکه هر موجودیت فقط یکبار شناسایی شود، کفایت می‌کند.

در متن زیر که به‌عنوان نمونه ارائه شده است:

TeliaSonera, the Nordic region’s largest telecoms operator, was formed in 2002 from the cross-border merger between Telia and Finland’s Sonera

مواردی که می‌تواند به عنوان اطلاعات استخراج شود، عبارت‌اند از:

FrameName: Merger
Company1: Telia
Company2: Sonera
New Company: TeliaSoner

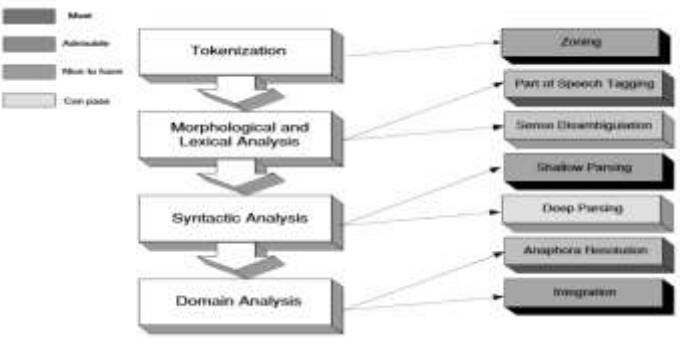
معماری نظام‌های استخراج اطلاعات

تصویر ۲ یک معماری تعمیم‌یافته را برای نظام استخراج اطلاعات نشان می‌دهد که برای فعالیت‌های پیش‌پردازش در داده‌کاوی مورد استفاده قرار می‌گیرد که در ادامه تشریح می‌شود.

نظام استخراج اطلاعات دارای سه تا چهار جزء اصلی است. نخستین جزء، ماژول توکنایزیشن یا تقسیم‌بندی و قطعه‌بندی[۳۰] و مشخص کردن واحدهای معنایی است که یک سند یا مدرک ورودی را به

27. Attributes
28. Facts
29. Event
30. Tokenization

سازه‌های اصلی آن تقسیم‌بندی می‌کند. سازه‌های اصلی معمولاً کلمات، جملات و بندها (پاراگراف‌ها) هستند و به‌ندرت ممکن است بلوک‌ها یا سازه‌هایی مانند بخش‌ها[۳۱] یا فصل‌ها[۳۲]را داشته باشیم. جزء دوم یک ماژول مورفولوژیکی، واژگانی، لکسیکال و نحوی[۳۳] است. این ماژول، بر فعالیت‌هایی مانند اختصاص برچسب‌های پارت‌آف‌اسپیچ‌تگینگ یا پی‌اُس[۳۴] (تحلیل اجزای کلام و جمله و گفتار و مشخص کردن نقش گرامری کلمات) به کلمات گوناگونِ اسناد و مدارک، ایجاد عبارت‌های اساسی مانند عبارت‌های اسمی و فعلی و رفع ابهام[۳۵] از کلمات و عبارت‌ها تمرکز می‌کند. جزء سوم، ماژولی برای تحلیل نحوی[۳۶] است که ارتباط بین بخش‌های گوناگون هر جمله را فراهم می‌کند. این فرایند به‌وسیله تحلیل کامل و عمیق فراهم می‌شود. چهارمین و متداول‌ترین جزء در هر نظام استخراج اطلاعات، با عنوان تحلیل دامنه[۳۷] (حوزه) نام‌گذاری می‌شود که عملکردی است که در آن، نظام، تمامی اطلاعات جمع‌آوری شده از اجزای پیشین را ترکیب می‌کند و یک چهارچوب کامل برای توصیف ارتباطات بین موجودیت‌ها فراهم می‌کند (فلدمن و سانجر، ۲۰۰۷). پس از شرح فعالیت استخراج اطلاعات، اکنون به تبیین فعالیت استخراج مفاهیم و تفاوت آن با استخراج اطلاعات پرداخته می‌شود.



تصویر ۲. معماری نظام استخراج اطلاعات

31. Sections
32. Chapters
33. Morphological & Lexical
34. Part Of Speech Tagging (POS)
35. Disambiguation
36. SyntacticAnalysis
37. DomainAnalysis

تعریف مفاهیم

مفاهیم، عبارت‌های کلیدی یا کلیدواژه‌های اصلی که در متن بحث شده است، نشان می‌دهند. آنها مجموعه‌ای از واژگان هستند که در متن اصلی وجود دارند و درحقیقت، خلاصه‌ای از متن اصلی را نمایش می‌دهند. از طرف دیگر مفاهیم، تعریفی مشابه عبارت‌های کلیدی دارند، با این تفاوت که نشان‌دهنده برخی مفاهیم و معانی اساسی موجود در متن هستند که از لغاتی تشکیل شداند که لزوماً در متن اصلی وجود ندارند. برای نمونه به متن برگرفته از روزنامه همشهری توجه کنید. جنگ جهانی دوم، جنگی فراگیر است که از سپتامبر ۱۹۳۹ آغاز شد و در اوت ۱۹۴۵ پایان یافت. این جنگ افزون‌بر اروپا در بخش‌های گسترده‌ای از قاره آسیا و آفریقا تأثیرات مخرب عمده‌ای بر جا گذاشت و کشورهای اسلامی از جمله ایران را درگیر خود کرد. علل اصلی جنگ جهانی دوم عبارت بود از اشتباهات عهدنامه ورسای که ظاهراً به جنگ جهانی اول پایان داد. همچنین پیامدهای بحران اقتصادی سال ۱۹۲۹ و از همه مهم‌تر رقابت سیاسی فاشیسم و دموکراسی‌های غربی و مارکسیسم. عامل اخیر چنان مؤثر بود که نبرد میان کشورهای درگیر به شکل بی‌سابقه‌ای، عموم مردم را به قلمرو جنگ کشاند، به‌طوری‌که در پایان جنگ تعداد کشته‌شدگان نظامی و غیرنظامی تقریباً با هم برابری می‌کرد. این جنگ به لحاظ گستردگی جغرافیایی و قدرت تخریب منابع انسانی و طبیعی بی‌همتا بوده است. آنچه که می‌توان به‌عنوان مفاهیم کلیدی از متن فوق استخراج کرد، قحطی، خشونت، ناامنی در جهان و ... است که این عبارت‌ها اگرچه در متن دیده نمی‌شوند اما می‌توان آنها را از متن درک کرد. استخراج مفاهیم به‌وسیله سیستم خودکاری ارائه می‌شود که توانایی استخراج مفاهیم کلیدی از متون را دارد (محمدی و بدیع، ۱۳۹۶). استخراج مفهوم، وابسته به زبان و زبان‌شناسی است و درتلاش برای کشف مفاهیم و معانی موجود در متن به جای شمارش واژه‌های پُرتکرار در متن است. در زبان‌شناسی، هر واژه دارای یک معنای ثابت بوده که تلقی فرد از آن یکسان است. این امر، به دلیل بافت زبان‌شناسی، نشانه‌شناسی و تاریخی منحصربه‌فرد یک واژه یا متن است که امکان مجزا شدن از معنای خود را نمی‌دهد. برخی از متخصصان، معنا را وابسته به سابقه ذهنی فرد می‌دانند و معتقدند که معنای یک واژه، به سابقه ذهنی فرد و میزان آشنایی او با آن بستگی دارد. معنا به موجودیت‌ها، اتفاق‌ها و واقعه‌هایی اشاره می‌کند که در ذهن کاربر قرار داشته و کاربر قصد دارد اطلاعاتی را درباره آن جستجو کند. استخراج مفاهیم، هم رایانه و هم انسان را که در یک همکاری دوسویه هستند، شامل می‌شود. پیش از پرداختن به استخراج مفاهیم و معانی در متن، به تعریف معنا می‌پردازیم.

چیستی معنا: خاستگاه و ریشه اصلی معنا، در زبان‌شناسی است، اما برخی از حوزه‌های علمی نیز همچون علم اطلاعات و دانش‌شناسی، ارتباط مستقیمی با معنا دارند و براساس هدف و نیاز خود، تعاریف مختلفی از معنا را ارائه می‌دهند. اما همه آنها به نوعی وام‌دار و وابسته به زبان‌شناسی هستند.

استخراج معنا از متن: فرض کنید که یک فرد دارای یک کیسه حاوی کلمات نامرتبی است که متن یک جمله یا یک بند (پاراگراف) را تشکیل می‌دهد. از «در کنار یکدیگر قرار دادن این کلمات»، معنا دریافت می‌شود. انسان‌ها دارای تجربه فرهنگی و زبان‌شناسی خاص خود هستند و به‌راحتی می‌توان‌ند با در کنار یکدیگر قرار دادن کلمات و برقراری رابطه بین آنها، معنا را کشف کنند، اما رایانه‌ها نمی‌توان‌ند این کار را انجام دهند و باید از پیش، مجموعه‌ای از اطلاعات برای آنها تعریف شود تا بتوانند بخشی از این کار را انجام دهند. استخراج مفهوم، یکی از مهم‌ترین ابزارهای طبقه‌بندی متن براساس مفاهیم است. استخراج معنا از یک متن، افزون‌بر اینکه جستجو را گسترش می‌دهد، تقویت‌کننده ارائه مدارک مرتبط‌تر با اصطلاح مورد جستجو نیز هست و دارای فواید دیگری نیز می‌باشد که شامل موارد ذیل است.

فواید استخراج معنا (مفهوم)

- ارائه مفاهیم مرتبط با اصطلاح مورد جستجو؛ این فایده، کاربر را با مفاهیم دیگری آشنا کرده، درنتیجه، دانش جدیدی را در اختیار وی قرار می‌دهد.
- ایجاد روابط بین مفاهیم و گسترش دامنه جستجو
- توسعه دانش زمینه‌ای
- تولید تفاسیر متون
- تعریف روابط بین متون
- طبقه‌بندی و دسته‌بندی متون

چالش‌های استخراج معنا از متن

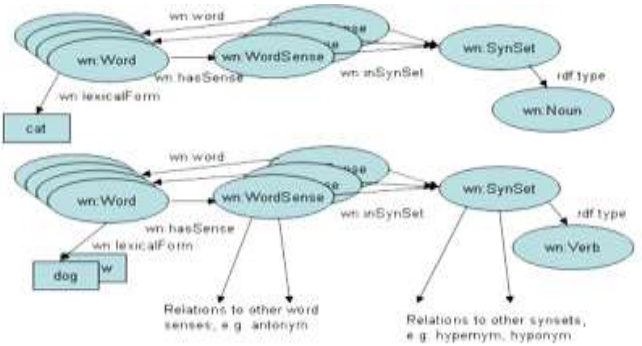
مفاهیم می‌توانند در قالب یک کلمه یا چندین کلمه بازگو شوند. در نظر گرفتن یک مفهوم در یک کلمه یا چندین کلمه، به زمینه موضوعی و میزان خاص و عام بودن آن بستگی دارد. برای نمونه، کلمه «شبکه»، یک مفهوم و «شبکه معنایی» مفهوم خاص‌تری دارد. بنابراین، یکی از چالش‌های استخراج مفاهیم، توجه به این مورد است. در یک زمینه موضوعی خاص نیز مفاهیم یک کلمه و چندکلمه‌ای به‌کار

می‌رود که برای گزینش آنها باید به میزان تکرارشان در متن توجه کرد (شیخ‌احمدی، ابوالحسنی، بخششی و خامفروش، ۱۳۸۶).

رویکردهای معمول استخراج مفاهیم از متن

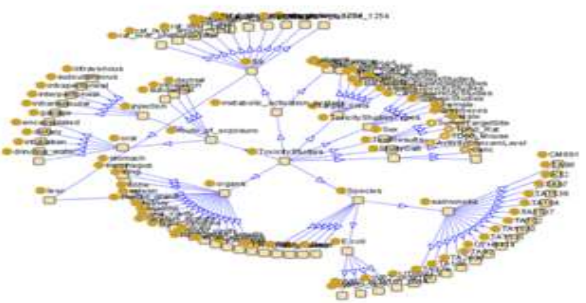
به‌طور کلی رویکردهای کمی برای استخراج مفاهیم (نکات کلیدی) به صورت خودکار وجود دارد که اغلب آنها نیز مبتنی‌بر روش‌های آماری (این رویکرد، به میزان تکرار یک کلمه در متن اشاره دارد) هستند. روش‌هایی که در سال‌های گذشته و اکنون در زمینه دریافت معنا از متن در رایانه به کار می‌روند، روش‌ها و رویکردهای مربوط به زبان‌شناسی و دانش‌محور است. این رویکردها، به ساختار دانشی متن و هستی-شناسی‌ها مربوط هستند که بیشتر بر زبان و ساختار آن تأکید می‌کنند که در ادامه به برخی از آنها پرداخته می‌شود.

شبکه‌های واژگانی: شبکه‌های واژگانی، ابزارهای ارزشمندی در پردازش زبان طبیعی هستند که برای نخستین بار برمبنای یافته‌های روان‌شناسی زبان ساخته شدند. این شبکه‌ها براساس روابط معنایی میان واژه‌ها شکل گرفته‌اند و تلاشی است در جهت بازنمایی آنچه در ذهن انسان‌ها از واژه‌ها و روابط آنها وجود دارد؛ بنابراین تلاش بر این است که حداکثر واژه‌های موجود در یک زبان را به صورت شبکه‌ای از روابط در خود بگنجانند. برای نمونه، وردنت، یک پایگاه داده لغوی بزرگ از لغات انگلیسی است. این بانک اطلاعاتی اسم‌ها، فعل‌ها، صفت‌ها و قیدها را به صورت مجموعه‌ای از لغات مترادف دسته‌بندی می‌کند که هر دسته، یک مفهوم مجزا را بیان می‌کند. مجموعه مترادف‌ها با استفاده از روابط معنایی– مفهومی و ارتباطات لغوی به یکدیگر پیوند داده شده‌اند. شبکه نهایی که شبکه‌ای از لغات و مفاهیم مرتبط از لحاظ معنایی است، می‌تواند توسط مرورگرها پیمایش شود. همچنین وردنت به صورت رایگان و برای عموم در دسترس و قابل بارگذاری است. ساختار وردنت، از آن یک ابزار مفید برای زبان‌شناسی محاسباتی و پردازش زبان طبیعی به وجود آورده است. وردنت، مشابه یک لغت‌نامه است که لغات را براساس معانی آنها دسته-بندی می‌کند. هر چند تفاوت‌های مهمی بین وردنت و دیگر لغت‌نامه‌ها وجود دارد. اوّل اینکه وردنت تنها شکل کلمات رشته‌هایی از حروف را پیوند نمی‌دهد، بلکه مفاهیم لغات را نیز مرتبط می‌سازد. درنتیجه لغاتی که در نزدیکی یکدیگر در شبکه یافت می‌شوند، قرابت معنایی نیز دارند. دومین تفاوت این است که وردنت روابط معنایی میان لغات را برچسب‌گذاری می‌کند. درحالی‌که دسته‌بندی‌های لغات در یک لغت‌نامه، از هیچگونه الگوی مشخصی جز مشابهت معنایی پیروی نمی‌کنند. تصویر زیر نمایی از وردنت است.



تصویر ۳. نمایی از وردنت

هستی‌شناسی: هستی‌شناسی الگویی انتزاعی از جهان واقع است که مفاهیم و روابط میان آن را در قلمروی موردبحث نمایش می‌دهد. هستی‌شناسی‌ها، پایگاه دانش مفهومی هستند. هستی‌شناسی‌ها به‌منزله ابزار بازنمایی و نمایش دانش در نظام‌های ذخیره و بازیابی، استفاده می‌شوند و آن را مجموعه‌ای از مفاهیم، ویژگی‌ها و روابط میان آن مفاهیم تعریف کرده‌اند. این تعریف در حوزه الگوسازی مفهومی، چندان جدید نیست. الگوهای موجودیت رابطه از دهه ۱۹۷۰ در پایگاه‌های اطلاعاتی استفاده می‌شود و در الگوهای گسترش‌یافته آن نیز چنین الگویی از مفاهیم، ویژگی‌ها و روابط قابل شناسایی است. اما دلیل این همه استقبال از هستی‌شناسی‌ها در این نکته نهفته است که هستی‌شناسی‌ها برخلاف الگوهای مفهومی پیش-گفته، استنتاج هوشمند را ممکن می‌سازند. تصویر زیر شمایی از هستی‌شناسی را نشان می‌دهد.



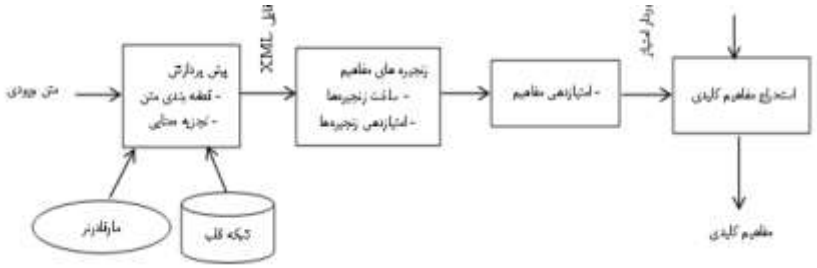
تصویر ۴: نمایی از هستی‌شناسی

رویکردی برای استخراج مفاهیم کلیدی با استفاده از شبکه قاب و زنجیره مفاهیم مبتنی‌بر پردازش زبان طبیعی: مراحل استخراج مفاهیم

گام اول در فرایند استخراج مفاهیم کلیدی، پیش‌پردازش است که شامل دو بخش قطعه‌بندی و سپس تفسیر معنایی متن ورودی است که در آن، مفهوم هر واژه، شناسایی می‌شود. با توجه به اینکه هر واژه ممکن است دارای چندین معنی باشد، یافتن نقش معنایی آن در جمله در این مرحله انجام می‌شود. شبکه قاب[۳۸]، ابزاری قدرتمند برای رسیدن به این هدف است. این شبکه، یک پایگاه داده لغوی در زبان انگلیسی است که قابلیت خوانایی توسط ماشین و انسان را دارد و مبتنی‌بر تفسیر نمونه‌هایی از چگونگی به‌کارگیری واژه‌ها در متن واقعی است. جملات تفسیرشده به صورت دستی در شبکه قاب وجود دارند که یک دادگان آموزش‌دیده منحصربه‌فرد را برای برچسب‌گذاری نقش‌های معنایی فراهم می‌کنند. آنچه‌که به‌عنوان خروجی این گام به‌دست می‌آید، متن برچسب‌گذاری‌شده‌ای است که در آن به هر واژه، مفهوم آن واژه نسبت داده شده است. گام بعدی، ساخت زنجیره‌های مفاهیم و سپس امتیازدهی به آنهاست. زنجیره مفاهیم، مفهومی مشابه زنجیره واژگان دارد، با این تفاوت که برای ساخت زنجیره واژگان از واژه‌های موجود در متن استفاده می‌شود. حال آنکه زنجیره مفاهیم با استفاده از مفاهیم (قاب‌های) موجود در متن ساخته می‌شوند. زنجیره مفاهیم دنباله‌ای از مفاهیم است که با یکدیگر ارتباط معنایی دارند و از طریق این روابط به یکدیگر متصل شده‌اند. زنجیره مفاهیم به صورت نموداری نمایش داده می‌شود که گره‌های آن مفاهیم (قاب‌ها) و یال‌های آن نشان‌دهنده ارتباط بین مفاهیم (قاب‌ها) است.

در گام پایانی مفاهیمی که اهمیت بیشتری دارند و عناوین اصلی متن ورودی را دربردارند، استخراج می‌شوند. برای رسیدن به این هدف، مفاهیم باید براساس معیارهای مناسبی انتخاب شوند. به همین منظور به هر مفهوم چهار امتیاز نسبت داده می‌شود که سه امتیاز آن با استفاده از زنجیره‌های مفاهیم به‌دست آمده‌اند. سپس حدود آستانه‌ای در نظر گرفته می‌شود و مفاهیمی که دارای امتیازی بیش از حد آستانه هستند، استخراج می‌شوند. برای ارزیابی سامانه نیز از مفاهیم کلیدی استخراج‌شده توسط انسان استفاده می‌شود. معماری نظام استخراج مفاهیم در تصویر ۵ نشان داده شده است. ورودی این سامانه، متن اصلی و خروجی آن، مجموعه‌ای از مفاهیم کلیدی است. همان‌طور که در تصویر مشاهده می‌شود، نظام استخراج دارای چهار مرحله است که خروجی هر مرحله، ورودی مرحله بعد می‌باشد. در ادامه این مراحل شرح داده می‌شوند.

38. Frament



تصویر ۵. معماری نظام استخراج مفاهیم

پیش‌پردازش متن: ورودی این مرحله متن اصلی است و خروجی آن یک فایل xml است که نشان‌دهنده تجزیه معنایی متن ورودی می‌باشد. این مرحله خود شامل دو زیربخش است، ابتدا متن ورودی قطعه‌بندی می‌شود و سپس عمل برچسب‌گذاری نقش‌های معنایی صورت می‌پذیرد.

۱-۱. قطعه‌بندی متن ورودی: در اولین مرحله متن ورودی باید با استفاده از روش‌های قطعه‌بندی، تقسیم شود. قطعه‌بندی متن، روند تقسیم متن ورودی به واحدهای معناداری نظیر کلمه، جمله یا موضوع است. قطعه‌کننده، ابزاری برای قطعه‌بندی متن است که در تلاش است متن را به قطعات معنادار تقسیم کند. ابزارهای مختلفی برای این هدف وجود دارند که هر یک از آنها از روش‌های قطعه‌بندی خاص خود استفاده می‌کنند. برای نمونه ابزار مارفادرنر[۳۹].

۲-۱. تجزیه معنایی: تجزیه معنایی متن به معنای یافتن ساختار معنایی آن است. استخراج نقش‌های معنایی یکی از گام‌های اصلی در بازنمایی معنی متن است. نقش‌های معنایی ارتباط معنایی بین فعل و آرگومان‌های آن را در جمله، مشخص می‌کنند. مشخص کردن ساختار معنایی متن، یکی از عملیات‌های اصلی و کلیدی در کاربردهای پردازش زبان طبیعی است (کاربردهایی نظیر استخراج اطلاعات، خلاصه-سازی، پرسش و پاسخ، شباهت معنایی، ترجمه ماشینی و ...). در سالیان اخیر بسیاری از پژوهشگران این موضوع را به‌عنوان یک مسئله برچسب‌گذاری مطرح کرده‌اند و از تکنیک‌های یادگیری ماشین برای ساخت چنین نظام‌هایی استفاده می‌کنند. برخی از این سیستم‌ها عمل آموزش را با استفاده از تفسیر شبکه قاب انجام می‌دهند و به‌طور خودکار تفسیرهایی برای متون ارائه می‌دهند. این روند را می‌توان برچسب‌گذاری خودکار نقش‌های معنایی نامید.

39. Marphadorner

۳-۱. ساخت زنجیره‌های مفاهیم: این مرحله شامل یک زنجیره مستقل از ساختار گرامری متن است. یک زنجیره مفاهیم، دنباله‌ای از مفاهیم مرتبط موجود در متن اصلی است. ساخت این زنجیره‌ها در ۳ مرحله انجام می‌شود.

الف) انتخاب ترم‌های کاندید: در این مرحله، قاب‌های موجود در متن به‌عنوان ترم‌های کاندید در نظر گرفته می‌شود. در واقع قاب، بیانگر مفهوم یک واژه در موقعیت خاصی از متن است. به همین دلیل زنجیره‌هایی که از قاب‌ها تشکیل شده‌اند، با عنوان زنجیره‌های مفاهیم نامگذاری شده‌اند.

ب) یافتن زنجیره مناسب: ارتباط و هماهنگی بین مفاهیم یکی از مسائل اساسی در ساخت زنجیره‌های مفاهیم است. سه نوع ارتباط بین مفاهیم تعریف شده است.

ارتباط بسیار قوی: که بین یک مفهوم و تکرارهای آن برقرار است. برای نمونه واحدهای لغوی plan.n و program.n هر دو قاب project را فراخوان می‌کنند. بنابراین اگر این دو واحد لغوی یا یکی از این دو به صورت مکرر در یک قطعه از متن وجود داشته باشند، باعث تکرار در فراخوانی قاب project می‌شوند.

ارتباط قوی: میان دو قاب که به‌وسیله یکی از انواع رابطه‌های قاب با قاب به یکدیگر متصل شده‌اند برقرار است. برای نمونه دو واحد لغوی children.n و young.a را در نظر بگیریم که به ترتیب قاب‌هایPeople- by- age و age را فراخوانی می‌کنند، این دو قاب با استفاده از یک رابطه قاب با قاب با یکدیگر در ارتباط‌اند.

ارتباط متوسط: میان دو قاب که به‌وسیله یک قاب دیگر با عنوان قاب واسطه به یکدیگر متصل شده‌اند برقرار است. برای نمونه دو واحد لغوی women.n و young.a که به ترتیب قاب‌های people و age را فراخوانی می‌کنند در نظر بگیرید. این دو قاب به‌وسیله قاب people- by- age به یکدیگر متصل شده‌اند که این قاب با عنوان قاب واسطه در نظر گرفته می‌شود.

ج) اضافه کردن قاب به زنجیره: هرگاه زنجیره مناسب قاب یافت شد، قاب به زنجیره اضافه می‌شود. به این ترتیب که بین قاب جدید و دیگر قاب‌های موجود در زنجیره که با آن ارتباط دارند، یال‌هایی متصل می‌شوند و زنجیره روزآمد می‌شود.

۴-۱. امتیازدهی به زنجیره‌های مفاهیم: در این مرحله زنجیره‌ها براساس معیارهای مختلف امتیازدهی می‌شوند. معیارهای متفاوتی برای امتیازدهی زنجیره‌ها وجود دارد از جمله تعداد عناصر (گره‌های) یک زنجیره، مجموع وزن یال‌های موجود در زنجیره و ... . برای نمونه براساس مجموع وزن یال‌های موجود

در زنجیره، هر زنجیره دارای دو امتیاز است: امتیاز ۱: تعداد عناصر (گره‌های) زنجیره؛ امتیاز ۲: مجموع فاصله‌های معنایی موجود در زنجیره (مجموع وزن یال‌های زنجیره).

۵-۱. امتیازدهی به مفاهیم: در این مرحله از شاخص‌های مختلفی برای امتیازدهی به مفاهیم استفاده می‌شود. نظیر بسامد مفهوم که عبارت است از «تعداد تکرارهای یک مفهوم در متن» و شاخص بیشترین وزن رابطه‌های مفهوم که برای محاسبه این ویژگی، مجموع وزن یال‌هایی را که به مفهوم متصل هستند در زنجیره‌های مختلف محاسبه کرده و بیشترین آنها به‌عنوان این شاخص در نظر گرفته می‌شود.

۶-۱. استخراج مفاهیم کلیدی: این مرحله، مرحله پایانی است. در این مرحله تعدادی از مفاهیم به عنوان مفهوم کلیدی استخراج می‌شوند. برای رسیدن به این هدف، سه حد آستانه در نظر گرفته شده است. اگر امتیاز یک مفهوم بیشتر از حد آستانه باشد، به‌عنوان مفهوم کلیدی استخراج می‌شود و درصورتی‌که امتیاز مفهوم کمتر از حد آستانه باشد، مفهوم کلیدی نخواهد بود.

سه حد آستانه با توجه به فرمول:

$$Score(concept) > Average(scores) + C \times StandardDeviation(Scores)$$

در نظر گرفته شده‌اند که ثابت C وابسته به تعداد کل مفاهیم موجود در متن یا تعداد مفاهیمی است که سامانه می‌تواند استخراج کند. سه حد آستانه با توجه به نمودار توزیع نرمال انتخاب می‌شوند. برای هر مفهوم یک بردار در نظر گرفته می‌شود و درنهایت مفاهیمی به‌عنوان مفهوم کلیدی استخراج می‌شوند که بردار مربوط به آنها بزرگتر از بردار حد آستانه باشد.

۷-۱. ارزیابی سیستم: پس از شناسایی و استخراج مفاهیم، ارزیابی سیستم به‌منظور حصول اطمینان از دقت و کارایی آن ضروری است. تورنی معتقد است که هر عبارتی به‌صورت بالقوه می‌تواند عبارت کلیدی باشد. اما تنها آن عبارت‌هایی کلیدی هستند که با عبارت‌های کلیدی انتساب‌شده توسط انسان مطابقت داشته باشند. بنابراین برای آزمایش و ارزیابی می‌بایست مفاهیم استخراج‌شده توسط سیستم (نرم‌افزار)، با مفاهیم استخراج‌شده توسط خبره مقایسه شود (محمدی و بدیع، ۱۳۹۶).

کاربردهای استخراج خودکار مفاهیم کلیدی

استخراج مفاهیم کلیدی در حل مسائلی نظیر کاربردهای تجارت الکترونیک، شاخص‌گذاری متون الکترونیکی، خلاصه‌سازی متون، موتورهای جستجو، خوشه‌بندی، دسته‌بندی، سامانه‌های پرسش‌وپاسخ، چکیده‌نویسی، بازیابی اطلاعات، ساخت کتابخانه‌های دیجیتال و .... می‌تواند مورد استفاده قرار گیرد. همچنین

یکی از کاربردهای بسیار بااهمیت استخراج مفاهیم، استخراج مفاهیم مربوط به هستی‌شناسی در دامنه‌های خاص است. هستی‌شناسی دو بخش دارد: مفاهیم و رابطه‌ها. تشخیص مفاهیم موجود در منابع داده برای ساخت یک هستی‌شناسی، یکی از کاربردهای سامانه خودکار استخراج مفاهیم کلیدی است. در سال ۲۰۱۶ پژوهشی انجام شده که مفاهیم هستی‌شناسی از میان چندین متن (تعداد متون محدودیت ندارد) استخراج کرده و تمامی متون در یک حوزه خاص قرار دارند (برای نمونه گزارش‌های هواشناسی، گزارش‌های ورزشی، پزشکی و ...). الگوریتم ارائه‌شده شامل ۳ مرحله است. در مرحله اول، عمل پیش‌پردازش متون انجام می‌شود که شامل قطعه‌بندی کردن کلمات است. سپس براساس الگوریتم ان‌گرام[۴۰]، مجموعه‌ای از عبارت‌ها کاندید تولید می‌شوند و در پایان با استفاده از قوانین آماری و زبان‌شناختی، مفاهیم را استخراج می‌کند. دو معیار اطلاعات متقابل و بسامد مستندات به‌عنوان قوانین آماری برای استخراج مفاهیم از بین عبارت‌های کاندید استفاده می‌شوند. با توجه به اینکه اطلاعات متقابل در تلاش برای انتخاب مفاهیم با بسامد پایین است، معیار بسامد مستندات در کنار آن، این نقص را برطرف می‌کند. حذف واژگان توقف نیز به‌عنوان قوانین زبان‌شناختی استفاده می‌شود. اگرچه در اینجا مفاهیم کلیدی مدنظر نیست، اما روال استخراج مفاهیم مشابه روند استخراج مفاهیم کلیدی است (مرادی، ۱۳۹۶؛ شیخ‌احمدی، ابوالحسنی، بخششی و خام‌فروش، ۱۳۸۶).

تفاوت‌های استخراج اطلاعات با استخراج مفاهیم

در استخراج اطلاعات، برخلاف داده، هیچ اطلاعات جدیدی ارائه نمی‌شود و اطلاعات موردنظر و مطلوب به‌ندرت با اطلاعات مشابه دیگری به‌طور همزمان وجود دارند. معمولاً در استخراج اطلاعات، با توجه به نیاز مطرح شده از سوی کاربر، مرتبط‌ترین متون و مستندات یا درواقع «کیسه کلمه» از میان دیگر مستندات یک مجموعه استخراج می‌شود. بازیابی اطلاعات، یافتن دانش نیست، بلکه تنها آن مستنداتی تحویل داده می‌شود که مرتبط‌تر به نیاز اطلاعاتی جستجوگر تشخیص داده می‌شود. این روش درواقع، هیچ دانش و حتی هیچ اطلاعاتی را به ارمغان نمی‌آورد. در مقابل، استخراج مفهوم، ربطی به جستجوی کلمات کلیدی در شبکه ندارد. این عمل در حوزه بازیابی اطلاعات گنجانده می‌شود. به‌عبارتی بازیابی اطلاعات جستجو، کاوش، طبقه‌بندی و فیلتر نمودن اطلاعاتی است که در حال حاضر شناخته شده‌اند و در متن قرار داده شده است. ولی در استخراج مفهوم، مجموعه‌ای از مستندات بررسی شده و اطلاعاتی که در هیچ یک از مستندات به صورت مجرد یا صریح وجود ندارد، استخراج می‌شود. به بیان دیگر می‌توان گفت، استخراج اطلاعات، اشاره به استخراج خودکار

40. N-grame

اطلاعات ساخت‌یافته همچون موجودیت‌ها، روابط بین موجودیت‌ها و موجودیت‌های توصیف ویژگی از منابع غیرساختاریافته دارد. این امر قالب‌های پرس‌وجو برای منابع غیرساختاریافته را نسبت به جستجوهای کلیدواژه‌ای مجرد، بسیار غنی‌تر می‌سازد. هنگامی‌که داده‌های ساختاریافته و غیرساختاریافته با هم موجود هستند، استخراج اطلاعات، یکپارچه‌سازی این دو نوع منابع و وضعیت جستجو بین آنها را امکان‌پذیر می‌سازد. در نقطه مقابل، استخراج مفهوم، وابسته به زبان و زبان‌شناسی است و سعی در کشف مفاهیم و معانی موجود در متن به جای شمارش واژه‌های پُرتکرار در متن دارد. همچنین می‌توان موارد زیر را به‌عنوان تفاوت‌های استخراج اطلاعات و استخراج مفاهیم برشمرد.

۱- در استخراج اطلاعات، موارد صریح مدنظر است، درحالی‌که در استخراج مفاهیم، مفاهیم ضمنی (غیرصریح) مدنظر قرار گرفته می‌شود.

۲- استخراج اطلاعات با شکل واژگان اما استخراج مفاهیم با معنای واژگان ارتباط دارد.

۳- خروجی استخراج اطلاعات، خلاصه‌ها، واژگان و کلیدواژه‌ها می‌باشند، درحالی‌که خروجی استخراج مفاهیم، هستی‌شناسی‌ها هستند.

۴- در هر دو الگوریتم‌ها و روش‌ها یکی هستند اما تکنیک‌ها یکی نیستند.

کاربرد استخراج اطلاعات و استخراج مفاهیم (داده‌کاوی) در سازمان‌های پلیسی و قضایی

امروزه کنترل و پیشگیری از جرایم و ایجاد و حفظ نظم، امنیت و آرامش در جامعه، از مهم‌ترین وظایف داعیه‌داران کشور در زمینه حفظ نظم، امنیت و آرامش (نیروی انتظامی، قوه قضائیه، وزارت اطلاعات، شورای امنیت ملی و دیگر سازمان‌های ذی‌ربط) به‌شمار می‌آید. جرایم ناهنجاری‌های اجتماعی هستند که جوامع به گونه‌های مختلف هزینه‌های زیادی برای آن می‌پردازند. شناسایی الگوهای جرم و کشف جرایم از دیرباز مورد توجه سازمان‌های قضایی و پلیسی بوده است. در گذشته از ابزارهای مختلفی برای تحلیل جرایم و کشف جرم و شناسایی روابط پیچیده میان جرم و مجرم استفاده می‌کردند. امروزه فناوری اطلاعات و تحولات الکترونیکی، باعث سهولت و تسریع در کشف و پیگیری جرایم و مجازات مجرمان و متهمان در جامعه امروزی شده است. سازمان پلیس نیز به‌عنوان یک سازمان پیشرو در عرصه فناوری اطلاعات با داشتن سامانه‌های اطلاعاتی رایانه‌ای برخط، با حجم زیادی از داده‌های اطلاعات عملیاتی در تمامی پلیس‌های تخصصی مواجه است. داده‌کاوی، ابزاری قدرتمند در تحلیل داده‌های جرم است که در کشف الگو و دانش از پایگاه داده جرم درراستای تصمیم‌گیری در سازمان‌های پلیسی به‌منظور کاهش فرصت ارتکاب جرم، کنترل جرایم و آسیب‌های اجتماعی نقش مؤثری دارد. می‌توان با داده‌کاوی بر داده‌های ذخیره‌شده و بهره‌گیری از نظر خبرگان، به دانش بسیار

باارزش در حوزه‌های مختلف پلیس دست یافت (مون، اِم‌سی‌کلاسکی‌الف و اِم‌سی‌کلاسکی‌ب[41]، ۲۰۱۰). امروزه داده‌کاوی و استخراج اطلاعات می‌توان‌د به‌عنوان ابزاری نوین به‌کار گرفته شود تا مشکلات و موضوعات شناسایی جرم را الگوسازی کند و به‌عنوان ابزاری قدرتمند برای سازمان‌های پلیس و جنایی مطرح شده و به‌کار گرفته می‌شود. تحلیل جرم عبارت است از به‌کارگیری یک شیوه نظام‌مند برای شناسایی، کشف و پیشگیری از جرایم. ماهیت پیچیده داده‌های مرتبط با جرم و بزهکاری و روابط نامحسوس میان این داده‌ها موجب مقبولیت روزافزون استفاده از دانش داده‌کاوی در میان جرم‌شناسان و تحلیل‌گران جرم شده است. درواقع دانش حاصل از اِعمال روش‌های داده‌کاوی در حوزه تحلیل جرم، بستر مناسبی برای پشتیبانی اطلاعاتی فرماندهان و مدیران به‌منظور انجام فعالیت‌های آتی پلیس فراهم می‌آورد. همچنان که کولین مک‌کیو[42] به‌عنوان مدیر پروژه واحد تحلیل جرم در سازمان پلیس ویرجینیا اظهار می‌کند: داده‌کاوی زمانی که در تحلیل جرم راه‌کنشی به‌کار برده می‌شود، ابزار اکتشاف دانش است که می‌تواند مجموعه داده‌های جامع را با سرعت بررسی کند و آرایه‌ای بیکران از متغیرها تهیه کند که این موضوع به مراتب برتر است از آنچه که یک تحلیل‌گر به‌تنهایی یک گروه تحلیلی یا گروه رزمی مشترک با دقت و درستی بررسی می‌کند (ابراهیمی، میرروشندل و آقایی، ۱۳۹۴).

کاربرد داده‌کاوی در حوزه‌های کاربردی پلیس

سه فعالیت مهم «شناسایی، پیشگیری و پیش‌بینی» همان‌گونه که در شکل زیر آمده است، در ارتباط با فعالیت‌های پلیس مطرح است (فیضی، لطفی و پیکری، ۱۳۹۵).

تصویر ۶. حوزه‌های کاربردی پلیس در رابطه با جرایم

41. Moon & McCluskey
42. Colleen McCue

برخی از روش‌ها و تکنیک‌های داده‌کاوی به‌منظور پیش‌بینی و پیشگیری از وقوع جرم

تکنیک‌ها، روش‌ها و الگوریتم‌های داده‌کاوی، ابزارها و راه‌های اجرای عملیات‌های داده‌کاوی‌اند که در ادامه به برخی از آنها اشاره می‌شود.

۱- قوانین تلازمی: قوانین تلازمی یکی از مهم‌ترین شیوه‌های داده‌کاوی است. شاید این روش را بتوان رایج‌ترین شکل روش‌های کشف الگو در نظام‌های یادگیری غیرنظارتی تلقی کرد. این شیوه داده‌کاوی بیشترین شباهت به رفتار مردم در هنگام شروع فراگیری داده‌کاوی دارد، یعنی «جستجوی طلا در پایگاه داده‌ای بسیار بزرگ». طلا در این مورد یعنی قاعده‌ای که چیزی درباره پایگاه داده می‌گوید که تاکنون درک‌پذیر نبود و اکنون که کشف شده، بسیار مورد توجه است. این روش‌ها تمامی الگوهای ممکن و دلخواه را درون پایگاه داده جستجو می‌کنند. قوانین کشف‌شده از یک سو می‌تواند نقطه قوت باشد، زیرا هیچ چیزی نیست که کاویده نشده باشد، از طرفی دیگر یک نقطه ضعف است، زیرا کاربر با انبوهی از قوانین روبرو خواهد بود که تحلیل کارایی آنها امری زمان‌بر و پرهزینه است.

۲- قوانین انجمنی: قوانین انجمنی، الگوهای مکرر موجود در داده‌ها هستند که می‌توانند هر گونه اختلاف را به‌عنوان یک نفوذ تشخیص دهند. برای اولین بار از تکنیک‌های کشف قوانین انجمنی فازی توسط بوکزاک و همکارانش در تحلیل داده‌های جنایی استفاده شد. استخراج قوانین انجمنی فازی در مطالعه جرم و جنایت بسیار مفید ارزیابی شده است (احمدی، منجمی و آیت؛ ۱۳۸۹)

۳- الگوریتم کای میانگین[۴۳]: الگوریتم کای میانگین از جمله ساده‌ترین و رایج‌ترین الگوریتم‌هایی است که برای خوشه‌بندی مجموعه داده‌های بزرگ به‌کار می‌رود. این الگوریتم که از معیار مربع خطا برای خوشه-بندی استفاده می‌کند، با یک تقسیم‌بندی اولیه و تصادفی، تخصیص نمونه‌ها به خوشه‌ها را آغاز می‌کند. براساس معیار شباهت بین نمونه‌ها و خوشه‌ها، تا زمانی که معیار همگرایی برآورده شود، فرایند تخصیص نمونه‌ها و محاسبه میانگین فاصله را تکرار می‌کند. طبیعی است درصورتی‌که نتوان هیچ نمونه‌ای را از خوشه‌ای جدا کرد و به خوشه‌ای دیگر تخصیص داد، به‌گونه‌ای‌که مربع خطا کاهش یابد، الگوریتم متوقف می‌شود.

۴- الگوریتم دسته‌های زمانی: الگوریتم دسته‌های زمانی، یک الگوریتم رگرسیون است که می‌تواند تعداد موردانتظار جرایم را برای یک سال پیش‌بینی کند.

43. K-Means

۵- الگوریتم درخت تصمیم: درخت تصمیم، درختی است که هر شاخه آن، یک انتخاب بین تعدادی از پیشنهادها را نشان می‌دهد، به‌نحوی‌که هر گره برگ، یک تصمیم را نمایش می‌دهد. درخت تصمیم به صورت عادی برای کسب اطلاعات به‌منظور تصمیم‌گیری استفاده می‌شود. درواقع درخت تصمیم، شیوه‌ای از ارائه یک سیستم است که تصمیم‌گیری‌های مهم را ساده و سیستم را به شیوه مناسبی تعریف می‌کند. با توجه به اینکه بیشتر سیستم‌های اجرایی و محاسباتی را می‌توان در قالب مجموعه‌ای از داده‌ها، ویژگی یا ویژگی‌ها و خروجی منطبق با آنها تعریف کرد، بنابراین می‌توان با استفاده از الگوریتم درخت تصمیم، ویژگی‌ها و خروجی‌ها را تحلیل کرد و سیستم را براساس این داده‌ها در قالب یک درخت تصمیم ارائه کرد. در درخت تصمیم، پیش‌بینی‌هایی مبنی‌بر برخی گرایش‌ها به سمت یک نتیجه خاص تهیه می‌شود. برای نمونه، یک درخت تصمیم پیش‌بینی می‌کند که آیا تبهکاران ویژه، چنین جرمی را انجام خواهند داد یا نه (فرهادی‌کالیانی و حسینی-همتی، ۱۳۹۶).

۶- الگوریتم شبکه بیضی: این الگوریتم احتمال مشروط بین ستون ورودی و قابل پیش‌بینی را محاسبه می‌کند و فرض می‌کند که ستون‌ها مستقل هستند.

۷- الگوریتم شبکه‌های عصبی: شبکه‌های عصبی مجموعه‌ای از اتصالات ورودی و خروجی است که هر اتصال دارای وزن ویژه خود می‌باشد. این الگوریتم شبیه مغز انسان است و شامل یادگیری الگوها از مجموعه داده‌ها برای پیش‌بینی جرم است. الگوریتم شبکه عصبی احتمالات را برای هر حالت ممکن از ویژگی‌های ورودی محاسبه می‌کند. الگوریتم شبکه عصبی برای تحلیل داده‌های ورودی پیچیده، بسیار مفید است و توانایی استنتاج معانی را از داده‌های مبهم و پیچیده دارد. برای نمونه، مأموران تحقیق و تفحص می‌توانند روش‌های شبکه عصبی را در استخراج موجودیت و پیش‌بینی جرم به‌کار برند (لی، کو و تِسای[۴۴]، ۲۰۱۰).

۸- شبکه‌های عصبی خودسازمانده[۴۵]: شبکه‌های عصبی خودسازمانده برای استخراج الگوهای داده‌های جرم به‌کار می‌روند. یکی از قابلیت‌های مهم این‌گونه شبکه‌ها، توانایی آنها در نگاشت داده‌های با ابعاد بالا در قالب داده‌هایی با ابعاد کمتر و در عین حال حفظ هم‌بندی[۴۶] میان آنهاست (کیوان‌پور، جاویده و ابراهیمی، ۱۳۸۸).

44. Li, Kuo & Tsai
45. Self Organizing Map (SOM)
46. Topology

۹- روش خوشه‌بندی: تکنیک خوشه‌بندی، اقلام داده را به داخل کلاس‌ها با ویژگی‌های مشابه و متناسب با حداکثر یا حداقل شباهت درون کلاسی طبقه‌بندی می‌کند، مانند شناسایی مظنونان که جرایم را به شیوه‌های مشابهی انجام می‌دهند یا تشخیص دادن گروه‌ها در میان دسته‌های مختلف جنایت‌کاران. این روش‌ها، مجموعه‌ای ازپیش‌تعریف‌شده برای اختصاص اقلام ندارند. به‌بیان‌دیگر، تکنیک‌های خوشه‌بندی، داده‌ها را براساس شباهت در یک کلاس قرار می‌دهند، داده‌های خوشه-بندی‌شده براساس اصل بیشترین شباهت بین اعضای هر خوشه و کمترین شباهت بین خوشه‌های مختلف گروه‌بندی می‌شود. از این‌رو می‌توان مظنونانی را که دارای حالت و ویژگی‌های مشابه هستند، شناسایی کرد یا نوع جنایت ارتکابی را از میان‌گروه‌های مختلف جرائم تشخیص داد. روش‌های خوشه‌بندی در پیوستگی و پیش‌بینی جرایم مؤثر هستند (وو[۴۷] و همکاران، ۲۰۰۸).

۱۰- نظریه تحلیل بقا: یکی از کارآمدترین روش‌های کشف پیش‌دستانه جرم و پیش‌بینی جرایم، نظریه تحلیل بقاست. درون‌مایه این نوع تحلیل بر این اصل استوار است که درصد چشمگیری از افرادی که به‌تازگی قربانی شده‌اند، با احتمال بالایی دوباره مورد جرم و جنایت قرار می‌گیرند. برای نمونه، این موضوع در مورد خانه‌هایی که در کشور انگلستان مورد سرقت قرار می‌گیرند، ثابت شده است. باید توجه داشت که هدف از این نوع تحلیل، برای نمونه تنها شناسایی خانه‌هایی که احتمالا در آینده مورد سرقت قرار می‌گیرند نیست، بلکه یکی از مهم‌ترین اهداف تحلیل بقا، شناسایی متغیرهای جرمی است که با وجود آنها احتمال بروز دوباره جرم بیشتر می‌شود. همچنین باید توجه داشت تمامی خانه‌هایی که به آنها دستبرد زده می‌شود، دوباره مورد سرقت قرار نمی‌گیرند، بنابراین شناسایی متغیرهایی که در جرایم ثانویه مشترک هستند، بسیار کاربردی به نظر می‌رسد.

۱۱- نظریه تحلیل ارتباطات: یکی دیگر از پرکاربردترین روش‌های پیش‌بینی جرم، استفاده از نظریه تحلیل ارتباطات است. این نوع تحلیل به شناسایی و کشف ارتباط بین عناصر مختلف درگیر در جرم می‌پردازد و مهمترین هدف آن، تشخیص اعضای تشکل‌ها و باندهای مجرمان است. این روش‌ها، سرکرده‌های باندهای مجرمان را نیز با دقت مناسبی شناسایی می‌کند. با تحلیل ارتباطات، هم می‌توان گروه‌های مجرمان را شناسایی کرد و هم با فرض وجود ارتباط بین دو جرم، از اطلاعات مربوط به یک جرم برای کشف و حل جرم دیگر نیز استفاده کرد (سیاح‌البرزی و رضایی، ۱۳۹۳؛ به نقل از لی، کو تسای، ۲۰۱۰)

47. Wu & et al

ترکیب قوانین تلازمی و الگوریتم کای میانگین و خوشه‌بندی برای کشف جرم

با به‌کارگیری قوانین تلازمی و کای میانگین و خوشه‌بندی، تلاش می‌شود تا الگوهای موردنیاز مدیران در سازمان‌های انتظامی، پلیسی و قضایی، کشف و در اختیار آنها قرار گیرد. الگوهای پنهان موجود در داده‌ها می‌تواند مدیران مربوط را در اخذ تصمیم‌گیری یاری نماید. به کمک ترکیبی از قوانین تلازمی و کای میانگین و خوشه‌بندی می‌توان به تشخیص و خوشه‌بندی مواردی چون استخراج الگوهای رفتاری مجرمان، سرقت، مناطق جرم‌خیز و خوشه‌بندی تبهکاران و متهمان پرداخت (اسکندری، علی‌زاده و کاظمی، ۱۳۹۱).

نرم‌افزارهای تحلیل جرم و داده‌کاوی

نرم‌افزارهای داده‌کاوی جرم نیز از دیگر ابزارهای داده‌کاوی هستند که در زیر به برخی از مهم‌ترین این نرم‌افزارها پرداخته می‌شود. بیشتر کاربران نرم‌افزارهای داده‌کاوی، با تفکر استفاده تجاری از این نرم‌افزارها خواهان استفاده از آن شده‌اند. نرم‌افزارهای داده‌کاوی معمولاً سه روش مختلف برای استفاده از داده‌کاوی به‌کار می‌برند: ۱. اکتشاف ۲. استفاده از مدل‌های پیشگویی و ۳. استفاده از تحلیل. اکتشاف، فرایند جستجو در داده‌هاست تا الگوهای مخفی موجود در داده‌ها را بدون هیچ تفکر ازپیش‌تعیین‌شده‌ای مشخص نماید. در نرم‌افزارهای داده‌کاوی مبتنی‌بر مدل‌های پیشگویی، الگوهایی که از یک بانک داده کشف می‌شوند، برای پیش‌بینی آینده به‌کار می‌روند. مدل‌های پیش‌بینی به کاربران اجازه می‌دهند تا داده‌های نامشخص را به‌کار ببرند و این مقادیر نامشخص توسط نرم‌افزار کشف شود. در مدل‌های تحلیلی نیز الگوهای یافت‌شده از داده‌ها برای تعیین مقادیر غیرعادی به‌کار می‌رود. برای تعیین مقادیر غیرعادی، ابتدا باید مقادیر عادی شناخته شود تا براساس آن، مقادیر غیرعادی و منحرف شناخته شوند (آناباتهوالا[48]، ۲۰۰۷).

چند نمونه از مهم‌ترین نرم‌افزارهای داده‌کاوی جرم

۱- نرم‌افزار کِرایم‌کانِکت[49]: این سیستم مبتنی‌بر وب است که برای اشتراک‌گذاری اطلاعات مربوط به جرم به‌کار می‌رود و به سازمان‌های قضایی و پلیسی کمک می‌کند تا اطلاعات خود، مانند

48. Annabathula
49. CrimeConnect

اشخاص مفقود، جرایم جنسی، ابلاغیه‌های رسمی و ... را در دنیای واقعی به اشتراک بگذارند. این سیستم، همچنین امکان جستجو در پایگاه داده جرم حرفه‌ای را برای افراد سازمان فراهم می‌کند تا به‌وسیله آن از سیستم‌های سایر حوزه‌های قضایی استفاده کنند و توانایی نیروهای پلیس را در حل جرایم افزایش دهند. درواقع نرم‌افزار کِرایم‌کانِکت، بستری را فراهم می‌کند تا حوزه‌های قضایی به‌سرعت و سهولت بتوانند اطلاعات مربوط به جرم را به داخل یک پایگاه داده امن وارد کرده و اطلاعات برخط همراه با داده از سایر منابع را به اشتراک بگذارند (اُتلی و ایوارت[۵۰]، ۲۰۰۳).

۲- نرم‌افزار کِرایم‌پوینت‌وب[۵۱]: برنامه کِرایم‌پوینت‌وب، یک راه‌حل نرم‌افزاری مبتنی‌بر وب است که در تسهیل به‌اشتراک‌گذاری اطلاعات، تحلیل و مدیریت اطلاعات برای نمایندگان اجرای قانون و امنیت عمومی کاربرد دارد (چونگ، چِن، چَنگ و چو[۵۲]، ۲۰۰۶).

۳- سیستم آسی[۵۳]: این سیستم بسیار مفید و کاربرپسند است که از همه جنبه‌های داده‌کاوی پشتیبانی می‌کند و مجموعه‌ای جامع از ابزارها برای انتخاب، دسترسی، دستکاری و آماده کردن داده فراهم می‌کند. این نرم‌افزار مبتنی‌بر جدیدترین فناوری‌ها و روش‌های استنتاجی است که مدل‌های پیشگویی را خودکار می‌سازد که در شکلی از درخت تصمیم نشان داده می‌شوند. آسی یک بسته جامع است که به فرد اجازه می‌دهد، وظایف داده‌کاوی پیشرفته را اجرا کند و سیستم‌های پشتیبان تصمیم قدرتمند را توسعه دهد (ماند، اِسرینیواس و مورتی[۵۴]، ۲۰۱۲).

۴- نرم‌افزار کارت[۵۵]: این نرم‌افزار ابزاری قدرتمند و مبتنی‌بر درخت تصمیم و بسیار جذاب است که به‌راحتی پایگاه‌های داده بزرگ و پیچیده را وارسی می‌کند و الگوها و وابستگی‌های مهم را جستجو کرده و جداسازی می‌کند. این دانش استخراج‌شده برای تولید مدل‌های پیشگویی استفاده می‌شود تا در اموری چون کشف ارتباطات از راه دور، سوءاستفاده از کارت‌های اعتباری و مدیریت خطرهای مالی به‌کار رود. این نرم‌افزار در مقایسه با انواع دیگر روش‌های تحلیل داده، یک پردازش اولیه بسیار خوب و کامل دارد، برای نمونه خروجی کارت (مقادیر پیش‌بینی‌شده) می‌تواند به‌عنوان یک

50. Oatley & Ewart
51. CrimePointWeb
52. Chung, Chen, Chang & Chou
53. AC
54. Mande, Srinivas & Murthy
55. CART

ورودی استفاده شود تا صحت و درستی پیش‌بینی روش‌های شبکه‌های عصبی و رگرسیون خطی را بهبود دهد (کاظمی و حسین‌پور، ۱۳۸۸).

۵- نرم‌افزار کِرایم‌اِستارسه[۵۶]: این نرم‌افزار یک برنامه آماری سه‌بُعدی برای تحلیل و شناسایی مکان‌های حوادث جرم است که توسط نِدلِوین[۵۷] و همکارش، براساس استاندارد مؤسسه ملی دادگستری توسعه‌یافته است. این برنامه مبتنی‌بر ویندوز با کاربرگه‌هایی مبتنی‌بر جی‌آی‌اِس[۵۸] است که ابزارهای آماری تکمیلی را برای کمک به نمایندگان اجرای قانون و پژوهشگران دادگستری جنایی در نگاشت جرم فراهم می‌کند. ورودی‌های برنامه، مکان‌های حوادث است. انواع آمارهای مقیاسی (سه‌بُعدی) را محاسبه می‌کند و نتایج را به صورت گرافیکی برای ویندوز و تحلیل سه‌بُعدی نمایش می‌دهد (مانیان، جمالو و بیدل، ۱۳۹۶).

تجربیات جهانی در سازمان‌های پلیسی و قضایی به‌منظور پیش‌بینی و پیشگیری از وقوع جرم

۱. داده‌کاوی و تحلیل حوادث ویژه: در ویرجینیای غربی حدود ۹۰۰ مرکز اجرای قانون وجود دارد که دَبلیووی‌اِس‌پی‌اِف‌اِل[۵۹]، یکی از این مراکز در جنوب ویرجینیای غربی است. این مرکز، امکاناتی را برای تحلیل نمونه‌های مربوط به حوادث ویژه فراهم کرده است. درواقع آزمایشگاه جنایی پلیس ایالت ویرجینیای غربی از یک ابزار نرم‌افزاری به نام فیلمز[۶۰] (سیستم مدیریت اطلاعات جنایی) استفاده کرده است. این ابزار با به‌کارگیری اصول آمار و الگوریتم‌های داده‌کاوی می‌توان‌د داده‌های جرم را تحلیل کند و نتایج منطقی را به‌دست آورد. این نرم‌افزار از سه قسمت تشکیل شده که عبارت‌اند از: آمار حوادث، داده‌کاوی و تصویرسازی داده. با کمک این ابزار، آمار حوادث به مراکز اجرای قانون ارائه می‌شود و آنان را در تصمیم‌گیری مربوط به بررسی و کنترل حوادث راهنمایی می‌کند. همچنین این ابزار می‌توان‌د در شناسایی الگوهای جرم در بین داده حوادث ذخیره شده در پایگاه داده فیلمز (جرم، مظنون و اطلاعات قربانی) به آزمایشگاه جنایی پلیس در ایالت ویرجینیای غربی کمک کند. این ابزار به دَبلیووی‌اِس‌پی‌اِف‌اِل و دیگر مراکز اجرای قانون در بهره‌برداری و پیش‌بینی فعالیت جنایی ممکن کمک خواهد کرد (احمدوند و آخوندزاده، ۱۳۸۹).

56. CrimeStat III
57. Ned Levine
58. GIS
59. WVSPFL
60. FIMS

۲. داده‌کاوی و بررسی صحنه وقوع جرم: هوش قانونی مانند اثر انگشت‌ها یا شناسایی دی‌اِن‌اِی به‌عنوان یک استاندارد روش قانونی برای تحقیق و کشف طیف وسیع از انواع جرم اهمیت دارد. پلیس نورتامپتونشیر[۶۱] از این داده‌های قانونی (اثر انگشت‌ها یا دی‌اِن‌اِی[۶۲]) و جرم برای اجرای طرح آزمایشی خود (بررسی صحنه وقوع جرم) استفاده کرد. مجموعه داده‌ها در هم ادغام شدند تا فعالیت فردی مرتبط با صحنه وقوع جرم را تولید کنند. از یک الگوریتم خوشه‌بندی یادگیری غیرنظارتی (کای میانگین) استفاده شد. یافته‌ها نشان داد که مأموران تحقیق می‌توانند براساس توانایی خود، دی‌اِن‌اِی یا اثر انگشت‌ها را از صحنه جرم جمع‌آوری کنند. همچنین توانایی آنها را در پیش‌بینی اینکه کدام یک از صحنه‌های جرم بهترین فرصت جمع‌آوری نمونه‌های قانونی را درحالی‌که با توانایی حقیقی آنها ارتباطی ندارد، نشان می‌دهد (بِل[۶۳]، ۲۰۰۶).

۳. داده‌کاوی و دوباره قربانی شدن: یک گروه تحقیقاتی متمرکز در دانشگاه ساندرلَند (مرکز سیستم-های قابل تطبیق)[۶۴]، از سوی سازمان تجارت و صنعت مأموریت داشت تا نرم‌افزاری با عنوان «نرم‌افزار هوشمند»[۶۵] برای تصمیم‌گیران آن سازمان پیاده‌سازی کند. نیروهای پلیس به بررسی پدیده‌های تکرار قربانی علاقه‌مند بودند. مفهوم دوباره‌قربانی‌شدن ابتدا توسط اِسپارکس[۶۶] بیان شد و بدین معناست که مکان‌هایی که در آن یک بار جرم اتفاق افتاده باشد، احتمال دارد که دوباره همان جرم در آنجا اتفاق بیفتد. برای نمونه احتمال سرقت دوباره پس از گذشت ۲۸ روز از اولین سرقت به صورت طبیعی (پس از ۶ ماه) کاهش می‌یابد. پس از آماده‌سازی نرم‌افزارهای ایوارت، اینگلیس و ویلبِرت[۶۷]، کم شدن فاصله زمانی بین دزدی‌های پی‌درپی و دوباره‌قربانی‌شدن در یک ملک را ثابت کردند (اُزکان[۶۸]، ۲۰۰۵).

۴. داده‌کاوی و حوادث تیراندازی: سازمان پلیس بین‌المللی وی‌آریچموند[۶۹]، ابزارهای اِس‌پی‌اِس‌اِس و آرتی‌آی[۷۰] را به‌کار برد تا حوادث تیراندازی اتفاقی را پیش‌بینی کند. بنابراین حوادث تیراندازی شب

61. Northamptonshire
62. DNA
63. Bell
64. Sunderland University (Compatible Systems Center)
65. Smart software
66. Sparks
67. Ewart, Inglis & Wilbert
68. Ozkan
69. VA Richmond
70. TI &SPSS

عید سال نو ۲۰۰۳ به میزان ۴۷ درصد بیش از سال پیش کاهش یافت. به‌کارگیری تحلیل‌های درخت تصمیم و روش‌های آنها به مأموران کمک کرد تا با سرعت بیشتری در برابر آن موقعیت در طول ۴۸ ساعت زمان بحرانی واکنش نشان دهد (اِکسو و بِراون[۷۱]، ۲۰۰۶).

۵. داده‌کاوی و سرقت‌های مسلحانه: بیشتر کارشناسان معتقدند هنگامی‌که سرقت‌های مسلحانه همراه با تهاجم خطرناک باشد، خسارت وارده به قربانی وخیم‌تر است. بنابراین، داده سرقت مسلحانه را در ریچموند ویرجینیا در یک دزدی احتمالی، آزمایش کردند تا خشونت‌های مسلحانه را محدود کنند. با استفاده از نرم‌افزار کِلِمِنتاین اِس‌پی‌اِس‌اِس، مدلی را توسعه دادند که عوامل وابسته به یک احتمال فزاینده از یک سرقت مربوط به حملات مسلحانه را مشخص کرد. سپس این نتایج را در نقشه‌ای برای استفاده توسط نیروهای گشت‌زنی و واحد راه‌کنشی گسترش دادند. نتایج نشان داد که مناطق با خطرهای بالا مشابه هم هستند، اما سرقت‌های مسلحانه به صورت یکسان در همه جا توزیع نشده است. مناطق با خطر بالا شامل مناطق جغرافیایی کوچک‌تری در شهر می‌شدند. با تعیین مناطق خطرخیز، امکان افزایش گشت‌زنی و گسترش واحدهای تاکتیکی در زمان بسیار کوتاه به صورت هدفمند، فراهم شد (کیوان‌پور، جاویده و ابراهیمی، ۱۳۸۸).

۶. داده‌کاوی و جرایم خشونت‌آمیز: درصورتی‌که فناوری هوش مصنوعی و الگوهای دقیق توسعه یابد، در پیش‌بینی اتفاقات آینده به‌کار گرفته می‌شود. آگاهی و بینش در مورد حوادثی که احتمالاً در آینده اتفاق خواهند افتاد، یک فرصت منحصربه‌فرد و حرفه‌ای را به مأموران اجرای قانون می‌دهد تا به صورت سریع واکنش نشان دهند. نشان دادن راه‌کنش‌های حرفه‌ای در مبارزه با جرایم از لحاظ اجتماعی بسیار مفید و مؤثر است. در گذشته این روش‌ها در مراکز تحقیقات دانشگاهی و بزرگترین آژانس‌های فدرال به صورت انحصاری وجود داشتند، اما اکنون این ابزارها در محیط میز کار رایانه‌های شخصی قابل دسترس هستند. سازمان پلیس ویرجینیا از این فناوری به‌منظور کنترل‌های محلی استفاده کرد و نظریه‌های خود را با اداره دادستان کل ایالات متحده در بخش شرقی از ویرجینیا با عنوان طرح تحقیقاتی پی‌اِس‌اِن[۷۲] و ارزیابی همکاران به اشتراک گذاشت. این طرح از داده‌کاوی و تحلیل‌های پیش‌گویانه استفاده کرد تا حوادث آینده را پیش‌بینی کنند و با به‌کارگیری و صف‌آرایی افراد پلیس، از جرایم خشونت‌آمیز جلوگیری کنند (آدِرلی[۷۳]، ۲۰۰۷)

71. Xue & Brown
72. PSN
73. Adderley

۷. داده‌کاوی و حملات تروریستی: پس از حمله تروریستی ۱۱ سپتامبر، سازمان‌های سیا، اِف‌بی‌آی و دیگر آژانس‌های فدرال تصمیم گرفتند تا اطلاعات داخلی و خارجی مربوط به حوزه امنیت را جمع‌آوری کنند تا بتوانند از حملات تروریستی جلوگیری کنند. این تلاش‌ها موجب ایجاد انگیزه در مقام‌های محلی شد تا به صورت دقیق‌تر جرایم قضایی حوزه خود را واپایش کنند. چالش اصلی تمام مجریان قانون و سازمان‌های گردآوری اطلاعات، دقت و مؤثر بودن میزان فزاینده تحلیل داده جرم است. برای نمونه حل کردن توطئه‌های پیچیده اغلب مشکل هستند، زیرا اطلاعات مظنونان ممکن است از نظر جغرافیایی و گستردگی در دوره‌های زمانی متفاوت باشد. همچنین تشخیص جرایم مجازی می‌تواند سخت باشد، زیرا شدآمد شبکه و تراکنش‌های برخط تکرارشونده، مقدار زیادی داده تولید می‌کند که تنها بخش کوچکی از فعالیت‌های غیرقانونی را تشریح می‌کند (ماند، اِسرینیواس و مورتی[۷۴]، ۲۰۱۲).

۸. داده‌کاوی و سرقت از منازل: طرح اُوِر[۷۵] در سال ۲۰۰۰، پلیس وِست مایدلَند[۷۶] را درگیر کرد و در مرکز سامانه‌های تطبیق‌پذیر و گروهی از روان‌شناسان دانشگاه ساندِرلَند قرار گرفت. هدف اصلی این طرح، کمک کردن به اداره و واپایش میزان جرم دزدی از منازل مسکونی با به‌کارگیری سامانه‌های پشتیبانی تصمیم، بود. پلیس به نرم‌افزاری نیاز داشت که به آنها در موارد زیر کمک کند: الف- هدف‌یابی منابع برای خط‌مشی‌های کشف و جلوگیری‌کننده به شکل بسیار مؤثر، ب- شناسایی داده‌های مهم برای اینکه در یک حادثه بتوان کارکنان را هدایت کرد و کارایی زمان را افزایش داد، ج- فراهم کردن اطلاعات درباره طراحی سیستم‌ها که این داده سخت (قانونی) و داده نرم (اطلاعات صحنه جرم) و اطلاعات هوشمندی پلیس را هماهنگ کند. تنوع روش‌های داده‌کاوی استفاده‌شده در این طرح (شبکه عصبی، خوشه‌بندی، تحلیل‌های بقا، شبکه‌های بیضی، دلایل موضوع‌محور، هستی‌شناسی و برنامه‌نویسی منطقی)، پلیس را در کشف مرتکبان دزدی از منازل مسکونی با نرخ اکتشاف ضعیف حمایت کرد (مالات‌هی و سانت‌هُش[۷۷]، ۲۰۱۱).

۹. داده‌کاوی و جرایم مجازی: اطلاعات مظنونان ممکن است از نظر جغرافیایی و گستردگی دوره‌های زمانی، متفاوت باشد. همچنین کشف جرائم فضای مجازی ممکن است مشکل باشد، زیرا شدآمد

74. Mande, Srinivas & Murthy
75. OVER
76. West Midland Police
77. Malathi & Santhosh

شلوغ شبکه و تراکنش‌های درون‌خطی تکرارشونده، داده‌های بسیاری تولید می‌کند که تنها بخش کوچکی از این فعالیت‌های غیرقانونی را تشریح می‌کند. داده‌کاوی ابزاری قدرتمند ارائه می‌دهد که به مأموران تحقیق جنایی که ممکن است فاقد آموزش باشند کمک می‌کند تا به‌عنوان تحلیل‌گران داده در اکتشاف پایگاه داده‌های بزرگ به‌سرعت و به‌طور مؤثر تحلیل نمایند. افزون‌بر این هزینه‌های نصب و راه‌اندازی (کارکرد) نرم‌افزار، اغلب کمتر از استخدام و آموزش کارکنان است. رایانه‌ها همچنین نسبت به نیروی انسانی تحقیق، به‌ویژه کسانی که ساعت‌های طولانی کار می‌کنند، کمتر در معرض اشتباه هستند. پژوهشگران دانشگاه آریزونا[۷۸] در همکاری با سازمان پلیس توسکان و فوئِنی[۷۹]، پس از سال ۱۹۹۷ مشغول به هدایت این موضوع هستند (لَک و رضایی‌نور، ۱۳۹۲؛ فرهادی‌کالیانی و حسینی‌همتی، ۱۳۹۶؛ به نقل از چونگ، چِن، چَنگ و چو، ۲۰۰۶)

نتیجه‌گیری

استخراج اطلاعات و استخراج مفهوم دو موضوع مهم در داده‌کاوی و تحلیل متن هستند. چنانچه که گفته شد، «استخراج اطلاعات»، بازیابی براساس کلیدواژه‌هاست، درحالی‌که «استخراج مفاهیم» با کلیدواژه‌ها ارتباط نداشته و با استخراج مفهوم و معنای استنباطی کاربر از متن، ارتباط دارد که این معنا ممکن است لزوماً در کلیدواژه‌ها نباشد. می‌توان گفت استخراج اطلاعات و استخراج مفاهیم یک چرخه است که این چرخه مدام غنی و پویا می‌شود. با توجه به کاربردهای ذکرشده استخراج اطلاعات و مفاهیم و داده‌کاوی و اهمیت تحلیل داده‌ها و تأکید نوشتار حاضر بر کاربرد داده‌کاوی در سازمان‌های پلیسی و قضایی به‌منظور کشف جرایم و این مطلب که یکی از پرکاربردترین کاربردهای داده‌کاوی در حوزه نیروی انتظامی است، همچنین حجم انبوه اطلاعات در نیروی انتظامی جمهوری اسلامی ایران و با توجه به قابلیت‌های روش‌های داده‌کاوی، در موارد زیر می‌توان این تکنیک را به کار برد:

۱- کشف ارتباط میان متغیرهایی ازقبیل سطح تحصیلات، جمعیت بیکار، جمعیت معتاد، پشتوانه مذهبی و سطح اقتصادی خانواده‌ها، نرخ جرم و جنایت، میزان شاخص آسیب‌های اجتماعی و دیگر متغیرهای کمّی یا کیفی مرتبط با شاخص‌های امنیتی در حوزه‌های استحفاظی کلانتری‌ها یا منطقه‌ای وسیع‌تر

78. Arizona
79. Tuscon & Phoeni

۲- تشخیص الگوهای بحران رفتاری در رابطه با اقدامات، فعالیت‌ها و معاملات
۳- خوشه‌بندی مناطق یا استان‌ها بر اساس میزان امنیت
۴- کشف رابطه میان انواع جرایم
۵- کشف رابطه میان شاخص‌های امنیتی با متغیرهای موجود
۶- دسته‌بندی کلانتری‌ها بر اساس نحوه عملکردشان
۷- خوشه‌بندی مجرمان
۸- پیش‌بینی وقوع انواع جرم
۹- دسته‌بندی جرایم
۱۰- امکان شناسایی، تحلیل و بررسی فرصت‌ها و تهدیدها
۱۱- ارائه راه‌کارهایی برای بهبود عملکرد در حوزه‌های مأموریتی و پشتیبانی (نجات و علی‌اکبری، ۱۳۸۸).

منابع:

منابع فارسی

ابراهیمی، مجیب؛ ابوالقاسم میرروشندل و جان‌احمد آقایی (۱۳۹۴). *جامعیت بخشی به مجموعه داده جرائم به‌منظور پیش‌بینی و شناسایی جرائم با استفاده از تکنیک‌های داده‌کاوی*. فصلنامه صنایع الکترونیک. ش (۴)۶ صص ۱۱-۵.

احمدوند، علی‌محمد؛ الهام آخوندزاده (۱۳۸۹). *داده‌کاوی، راهی نوین در تحقق پلیس پاسخگو*. فصلنامه نظارت و بازرسی. ش (۱۱)۴. صص ۵۴-۳۴.

احمدی، محمدحسین؛ امیرحسین منجمی و سعید آیت (۱۳۸۹). *دسته‌بندی متون فارسی با استفاده از قواعد انجمنی*. چهارمین کنفرانس داده‌کاوی ایران. تهران: دانشگاه صنعتی شریف.

اسکندری، حمیدرضا؛ سمیه علیزاده و پروانه کاظمی (۱۳۹۱). *کاربرد داده‌کاوی در شناسایی و کشف الگوهای پنهان جرم سرقت*. *فصلنامه نظم و امنیت انتظامی*. ش (۴)۴. صص ۵۶-۳۵.

اسماعیلی، مهدی (۱۳۹۶). *مفاهیم و تکنیک‌های داده‌کاوی*. تهران: انتشارات نیاز دانش.

ترکیان، ایوب (۱۳۹۷). متن‌کاوی: نگرش یادگیری ماشین. تهران: نیاز دانش.

تیمورپور، بابک؛ سمیه علیزاده و مهدی غضنفری (۱۳۹۵). داده‌کاوی و کشف دانش. تهران: دانشگاه علم و صنعت ایران.

حاصلی، داوود؛ ملوک‌السادات حسینی‌بهشتی و سمیه پاک‌نهاد (۱۳۹۵). استخراج اطلاعات: روش‌ها و کاربردها. اولین کنفرانس بین‌المللی بازیابی تعاملی اطلاعات. قابل دسترس در: https://www.civilica.com/

حریری، نجلا (۱۳۹۰). نظام‌های بازیابی اطلاعات متنی. تهران: چاپار.

رشادت، وحیده؛ مریم حورعلی (۱۳۹۳). «*مروری بر روش‌های استخراج رابطه در یادگیری هستان نگار و استخراج اطلاعات*». همایش ملی مهندسی رایانه و مدیریت فناوری اطلاعات.

زاهدی احمدسرایی، عباس؛ فرشید مهردوست (۱۳۹۳). *مروری بر الگوریتم‌های داده‌کاوی و روش‌های استخراج اطلاعات با معرفی کاربرد داده‌کاوی در امور بانکی*. اولین همایش ملی پژوهش‌های مهندسی رایانه.

سیاح‌البرزی، هدایت؛ محسن رضایی (۱۳۹۳). *نقش رویکردهای نظارتی و راهبردیِ سیاست کیفری بر مجرمان سابقه‌دار در پیشگیری از تکرار جرم در ایران*. فصلنامه مطالعات پیشگیری از جرم. ش (۲۸)۸. صص ۹۹-۱۲۲.

شیخ‌احمدی، سیدامیر؛ حسن ابوالحسنی، هاشم بخششی و کیهان خام‌فروش (۱۳۸۶). *یک روش پیشنهادی برای استخراج مفاهیم دامنه*. سیزدهمین کنفرانس سالانه انجمن کامپیوتر ایران.

فرهادی‌کالیانی، مسلم؛ سامان حسینی‌همتی. (۱۳۹۶). *پیش‌بینی جرائم سایبری با بهره‌گیری از روش‌های داده‌کاوی و ارائه یک الگوریتم بهینه: مطالعه موردی پلیس فتا استان کرمانشاه*. چهارمین کنفرانس ملی مهندسی برق، کامپیوتر و فناوری اطلاعات.

فیضی، کامران؛ رسول لطفیو ناصر پیکری. (۱۳۹۵). *کاربرد داده‌کاوی در استفاده بهینه از بانک‌های اطلاعاتی پلیس: راهبردهای نوین*. فصلنامه مطالعات راهبردی ناجا. ش (۲)۱. صص ۱۷۶-۱۵۹.

کاظمی، پروانه ؛ جواد حسین‌پور. (۱۳۸۸). *کاربرد داده‌کاوی در سازمان پلیسی و قضایی به‌منظور شناسایی الگوهای جرم و کشف جرایم*. فصلنامه کارآگاه. ش ۸ صص ۶۳-۳۲.

کرمی، راضیه؛ ملیحه سادات ملک‌جعفریان. (۱۳۹۳). *اهمیت پردازش داده‌ها*. فصلنامه آمار. ش (۴)۲. صص ۳۶-۳۴.

کیوان‌پور، محمدرضا؛ مصطفی جاویده و محمدرضا ابراهیمی. (۱۳۸۸). *تحلیل رایانه‌ای جرم با بهره‌گیری از روش‌های هوش مصنوعی و داده‌کاوی کشف پیش‌دستانه جرم*. فصلنامه کارآگاه. ش (۷)۲. صص ۹۸-۱۱۷.

لَک، بهزاد؛ جلال رضایی‌نور. (۱۳۹۲). *تحلیل و کشف جرم از طریق کاوش متون در فضای مجازی*. فصلنامه توسعه مدیریت منابع انسانی و پشتیبانی. ش (۲۷)۸. صص ۱۸۲-۱۵۹.

مانیان، امیر؛ محمد جمالو و معصومه بیدل. (۱۳۹۵). *طراحی الگوی داده‌کاوی پیشنهادی به‌منظور شناسایی مجرمان*. فصلنامه انتظام اجتماعی. ش (۳)۸. صص ۱۲۸-۱۰۹.

محمدی، سودابه؛ کامبیز بدیع. (۱۳۹۶). *استخراج مفاهیم کلیدی با استفاده از شبکه قاب و زنجیره مفاهیم*. نشریه مهندسی برق و مهندسی کامپیوتر ایران. ش (۱)۱۵. صص ۶۴-۷۲.

مرادی، خدیجه. (۱۳۹۶). *استخراج مفهوم در داده‌کاوی*. نشریه رهاورد نور.

نجات، امیررضا؛ آرش علی‌اکبری. (۱۳۸۸). *داده‌کاوی، راهی به سوی ناشناخته‌ها*. فصلنامه توسعه سازمانی پلیس. ش (۱۸)۵. صص ۶۹-۵۳.

منابع لاتین

Adderley, R & Musgrove, P.B. (2001). *General review of Police crime recording and investigation systems: A user's view.* Policing: An International Journal of Police Strategies and Management, 24(1).

Annabathula, R (2007). *A web-based tool for analysis of crime laboratory data.* Master of Science Thesis, West Virgina University.

Bell, Chris. (2006). *Concepts and possibilities in forensic intelligence.* Forensic Science International, 162(1-3): 38-43.

Chung, Wingyan; Chen, Hsinchun; Chang, Weiping & Chou, Shihchieh. (2006). *Fighting cybercrime: a review and the Taiwan experience.* Decision Support Systems, 41(3): 669-682.

Feldman, Ronen, Sanger, James. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge University Press, New York.

Li, Sheng-Tun, Kuo, Shu-Ching & Tsai, Fu-Ching. (2010). *An intelligent decision-support model using FSOM and rule extraction for crime prevention.* Expert Systems with Applications, 37 (10): 7108-7119.

Malathi, A & Baboo, S. Santhosh. (2011). *An Enhanced Algorithm to Predict a Future Crime using Data Mining.* International Journal of Computer Applications, 21(1): 1-6.

Mande, Uttam; Srinivas, Y & Murthy, J. V. R. (2012). *An Intelligent analsis of crime data using data mining & auto correlation models.* International journal of engineering research and applications, 2(4), : 149-153.

Moon, Byongook; McCluskey, John D, (a) & McCluskey (b), Cynthia Perez. (2010). *A general theory of crime and computer crime: An empirical test.* Journal of Criminal Justice, 38 (4): 767-772.

Oatley, Giles. C & Ewart, Brian. W. (2003). *Crimes analysis software: 'pins in maps', clustering and Bayes net prediction.* Expert Systems with Applications, 25(4): 569-588.

Ozkan, Kadir. (2005). *Managing data mining at digital crime investigation.* Forensic Science International, 146: 37-38.

Wu, Xindong; Kumar, Vipin; Quinlan, J. Ross; Ghosh, Joydeep; Yang, Qiang; Motoda, Hiroshi; McLachlan, Geoffrey J.; Ng, Angus; Liu, Bing; Yu, Philip S.; Zhou, Zhi-Hua; Steinbach, Michael; Hand, David J. & Steinberg, Dan. (2008). *Top 10 algorithms in data mining.* Knowledge and Information Systems, 14: 1-37.

Xue, Yifei & Brown, Donald E. (2006). *Spatial analysis with preference specification of latent decision makers for criminal event prediction.* Decision Support Systems, 41(3): 560-573.