# استخراج مفهوم در داده‌کاوی


**خدیجه مرادی***
**atefehmorady@gmail.com**

**مقدمه**

استفاده از رایانه در امور مختلف، باعث شده تا داده‌های بسیاری با سرعت‌های زیاد در پایگاه داده‌ها انباشته و ذخیره شوند. پردازش این داده‌های حجیم، خارج از توان انسان است. تلاش‌های فراوانی تاکنون انجام شده است تا نرم‌افزارها و سخت‌افزارها توسعه پیدا کنند و تولید، ذخیره و انتقال داده‌ها انجام گردد؛ اما تجزیه و تحلیل این حجم از داده‌ها توسط رایانه‌ها، بعد از ذخیره و پردازش، تاکنون انجام نشده است. داده‌ها در عصر حاضر، قلب تپنده هر سازمان را تشکیل می‌دهند و هر روز به میزان داده‌ها در سیستم‌های اطلاعاتی افزوده می‌شود. در واقع، سازمان‌ها در اطلاعات غرق شده‌اند؛ درحالی‌که تشنه دانش هستند. این امر، نشانگر آن است که سازمان‌ها نتوانسته‌اند از دانش درون داده‌ها به نحو مناسب استفاده نمایند. در درون حجم عظیمی از داده‌ها، الگوها و روابط بسیاری میان پارامترهای مختلف به صورت پنهان باقی می‌ماند که برای برنامه‌ریزی‌های استراتژیک و طولانی‌مدت می‌تواند حیاتی باشد. بنابراین، نیاز به ابزاری است تا داده‌ها را به‌گونه‌ای پردازش کند تا دانش حاصل از آن را در اختیار تصمیم‌گیران سازمان قرار دهد. یکی از راهکارهایی که امروزه در این زمینه ایجاد و در حال گسترش است، داده‌کاوی(۱) است. داده‌کاوی، فرآیند کشف دانش پنهان درون داده‌هاست که با توصیف، تشریح، پیش‌بینی و کنترل پدیده‌های گوناگون پیرامونی، دارای کاربرد بسیار وسیعی در حوزه‌های مختلف است؛ به‌گونه‌ای‌که مرز و محدودیتی برای کاربرد آن در نظر گرفته نشده و زمینه‌های کاربردی آن را از ذرات کف اقیانوس تا اعماق فضا می‌دانند (شهرابی، ۱۳۸۶).

تعریف داده، اطلاعات و دانش کار، مشکل است

* دانشجوی دکترای علم اطلاعات و دانش‌شناسی دانشگاه الزهراء

و تنها از دیدگاه کاربران و استفاده‌کنندگان می‌توان تمایزی بین آنها ایجاد نمود و آنها را از یکدیگر تشخیص داد. از دیدگاه صاحب‌نظران، داده به واقعیت‌های ساخت‌نیافته و بدون ساختار اطلاق می‌شود که به‌تنهایی معنا ندارد؛ به بیانی دیگر، داده‌ها حقایق خام هستند و در پایگاه‌های داده ذخیره و مدیریت می‌شوند و تا زمانی که پردازش نشوند، هیچ برداشتی از آنها صورت نمی‌گیرد. اطلاعات، داده‌های خلاصه‌ای هستند که گروه‌بندی، ذخیره، پالایش و سازماندهی شده‌اند تا بتوانند معنادار شوند. اطلاعات زمانی ارزش پیدا می‌کند که برای یک فرد، هدف و در زمان خاص گردآوری و آماده‌سازی شود. دانش، عبارت است از اطلاعات دسته‌بندی‌شده و مرتبط که کاربرد اجرایی و عملی یافته‌اند (داونپورت(۲) ، ۱۹۹۸). دانش، عبارت است از مجموعه باورها، مهارت‌ها، شناخت‌ها، تئوری‌ها، مقررات و اقدامات عملی که سازمان و افراد آن را در اختیار دارند و برای اتخاذ تصمیمات و حل مسائل مختلف از آن استفاده می‌کنند. دانش، آن بخشی از اطلاعات است که در عمل برای اخذ تصمیمات و انجام اقدامات به‌کارمی‌رود (باقری و سلاجقه، ۱۳۸۹).


**شکل ۱: هرم داده تا خِرد**

## کشف دانش(۳)

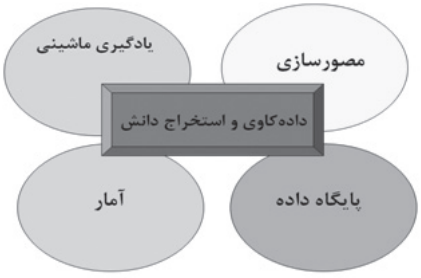
فرایند استخراج دانش مفید از میان انبوهی از داده‌ها، «کشف دانش» نامیده می‌شود. کشف دانش، مربوط به کل فرآیند استخراج دانش است؛ از جمله: چگونه داده‌ها ذخیره و قابل دسترس می‌شوند، چگونه با استفاده از الگوریتم‌های کارآمد و مقیاس‌پذیر به تجزیه و تحلیل مجموعه داده‌های عظیم می‌پردازد، چگونه به تفسیر و مصورسازی نتایج می‌پردازد و چگونه به مدل‌سازی و پشتیبانی تعاملات میان انسان و ماشین می‌پردازد. همچنین، مرتبط با پشتیبانی برای یادگیری و تحلیل در برخی حوزه‌های کاربردی است. داده‌کاوی و استخراج دانش از پایگاه داده‌ها، از موضوعاتی است که هم‌زمان با تولد پایگاه‌های اطلاعاتی ایجاد شد. پژوهشگران و متخصصان رایانه، هوش مصنوعی، یادگیری ماشینی و غیره، از اوایل دهه ۲۰میلادی، به پژوهش در این زمینه پرداختند )تاج‌الدینی، موسوی و دلیری، ۱۳۸۸). عبارت «کشف دانش در پایگاه داده‌ها» منسوب به کارگاهی آموزشی در سال ۱۹۸۹ است. این عبارت، به این معناست که نتایج نهایی از تجسس داده‌ها، باید درکشف دانش قابل استفاده باشد )فایاد(۴) ، ۱۹۹۶).

کشف دانش، به دو روش انجام می‌شود:

۱.داده‌کاوی یا استخراج دانش از پایگاه داده‌ها؛

۲.متن‌کاوی یا استخراج دانش از متن (هوتو(۵) ، ۲۰۰۵).

داده‌کاوی و استخراج دانش از پایگاه داده‌ها هم‌زمان با تولد پایگاه اطلاعاتی ایجاد گردید و توسط متخصصان آمار، یادگیری ماشینی، تحلیل‌گران داده، پژوهشگران هوش مصنوعی و مصورسازی، در اوایل دهه ۸۰ برای اوّلین‌بار به‌کار گرفته شد. بعدها با افزایش رشد فناوری اطلاعات، استفاده از آن در بسیاری از علوم متداول گشت. اوّلین گزارش رسمی از داده‌کاوی را به لوول در سال ۱۹۸۳ نسبت می‌دهند؛ اما به طور کلی، پژوهش‌های جدی در این زمینه، از دهه ۹۰ آغاز گردیده است (تاج‌الدینی و موسوی، ۱۳۸۸).




**شکل ۲: عوامل موثر بر استخراج دانش**

## داده‌کاوی

داده‌کاوی یا استخراج دانش از پایگاه داده‌ها(۶)، ترجمه اصطلاح data mining است. نگاهی

**داده‌کاوی، فرآیند کشف دانش پنهان درون داده‌هاست که با توصیف، تشریح، پیش‌بینی و کنترل پدیده‌های گوناگون پیرامونی، دارای کاربرد بسیار وسیعی در حوزه‌های مختلف است؛ به‌گونه‌ای‌که مرز و محدودیتی برای کاربرد آن در نظر گرفته نشده و زمینه‌های کاربردی آن را از ذرات کف اقیانوس تا اعماق فضا می‌دانند**

به این ترجمه تحت اللفظی تقریباً مفهوم آن را نشان می‌دهد. Mine، به معنای استخراج از منابع نهفته و با ارزش زمین است و ترکیب داده با کلمه mine، جست‌وجوی عمیق برای کشف و پیداکردن اطلاعات اضافی مفید را که قبلاً نهفته بودند، از داده‌های حجیم پیشنهاد می‌دهد (برسون(۷)، ۲۰۰۴). داده‌کاوی، نام خود را از شباهت‌هایی که بین دو حوزه جست‌وجوی اطلاعات مرتبط و ارزشمند در پایگاه‌های داده‌ای بزرگ و کندوکاو سنگ‌ها در معادن برای استخراج رگه‌هایی از سنگ‌های معدنی ارزشمند به عاریت گرفته است و این دقیقاً همان چیزی است که داده‌کاوی دنبال می‌کند و هدف آن است. داده‌کاوی، به دنبال استخراج اطلاعات ارزشمند از یک پایگاه اطلاعاتی بزرگ است (خاصه، ۱۳۸۹). برخی، داده‌کاوی را معادل استخراج دانش، وارسی داده‌ها و لایروبی داده‌ها می‌دانند.

از داده‌کاوی تعاریف مختلفی در متون مختلف ارائه شده است؛ برخی از تعاریف، داده‌کاوی را به عنوان ابزاری برای ارتباط کاربر با داده‌های عظیم معرفی می‌کند و برخی دیگر، به کاوش داده‌ها و ارائه تعاریف دقیق‌تری از داده‌کاوی پرداخته‌اند که در زیر به شماری از آنها اشاره می‌گردد:

- داده‌کاوی، فرایند کشف دانش پنهان موجود در داده‌هاست (شهرابی، ۱۳۸۶)؛
- داده‌کاوی، یک فرآیند شناخت الگوهای معتبر، جدید، ذاتاً مفید و قابل فهم از داده‌هاست؛
- کشف دانش از پایگاه داده‌ها؛
- فرآیند کشف الگوهای مفید از داده‌ها؛
- فرآیند انتخاب، کاوش و مدل‌بندی داده‌های حجیم، جهت کشف روابط نهفته با هدف به‌دست‌آوردن نتایج واضح و مفید، برای مالک پایگاه داده‌ها (گیورودی(۸)، ۲۰۰۱).


متداول‌ترین تعریف از داده‌کاوی که در بیشتر متون و مراجع به کار برده شده، عبارت است از: «استخراج اطلاعات و دانش و کشف الگوهای پنهان از یک پایگاه داده بسیار بزرگ و پیچیده.» داده‌کاوی کمک می‌کند تا سازمان‌ها با کاوش بر روی داده‌های یک سیستم، الگوها، روندها و رفتارهای آینده را کشف و پیش‌بینی کنند و بهتر تصمیم بگیرند. داده‌کاوی با استفاده از تحلیل وقایع گذشته، یک تحلیل خودکار و پیش‌بینانه ارائه می‌نماید و به سؤالاتی جواب می‌دهد که پاسخ آنها در گذشته ممکن نبوده و یا به زمان بسیاری نیاز داشت. ابزارهای داده‌کاوی، الگوهای پنهانی را کشف و پیش‌بینی می‌کنند که متخصصان ممکن است به دلیل اینکه این اطلاعات و الگوها خارج از انتظار آنها باشد، آنها را مدنظر قرار ندهند و به آنها دست نیابند (بری(۹)، ۱۹۹۷).

### مراحل داده‌کاوی

داده‌کاوی، شامل مراحل زیر است:

۱. شناخت: اوّلین مرحله این است که ما حوزه موضوعی و کاربری مورد نظری را که قرار است داده‌کاوی در آن انجام شود، بشناسیم. نیازها و هدف‌های کاربران را بررسی کرده، دانشی در آن باره کسب نماییم.

۲. داده‌های هدف: مجموعه هدف را انتخاب کنیم و داده‌های مورد نظر خود را انتخاب نماییم.

۳. پاک‌سازی داده‌ها(۱۰): داده‌های اضافی و

نامربوط از مجموعه‌داده‌ها حذف می‌شوند.

۴. تبدیل و تغییر شکل داده‌ها: داده‌های انتخاب‌شده به صورتی متناسب جهت انجام فرایند داده‌کاوی تبدیل می‌شوند.

۵. داده‌کاوی: این مرحله، سخت‌ترین مرحله است و در آن از تکنیک‌های هوشمند همچون: قوانین انجمنی، خوشه‌بندی، طبقه‌بندی و... برای استخراج الگوهای مفید استفاده می‌شود.

۶. ارزیابی و تفسیر الگوها(۱۱)؛

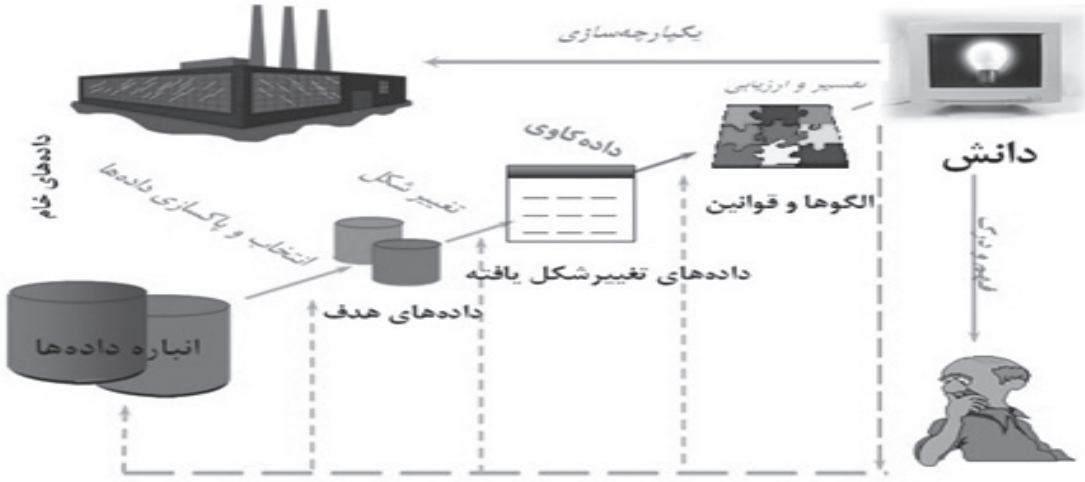
۷. دانش.

• عدم نیاز به اعتبارسنجی نتایج محصول استخراج مدل از داده‌هاست.

• اشکالات معمول حاصل از نمونه‌گیری آماری، در داده‌کاوی وارد نمی‌شود. حجم داده در داده‌کاوی، بستگی به موضوع داده‌کاوی دارد.

• داده‌کاوی، نیازمند به فرضیه نیست و ممکن است برای اثبات فرضیه از داده‌کاوی استفاده شود.

• در داده‌کاوی، آنالیزهای واقعی صورت می‌گیرد و به توزیع داده‌ای وابسته نیست. از دیدگاه




**شکل ۳: فرایند داده‌کاوی**

**اهمیت و لزوم داده‌کاوی**

در این قسمت، به این سؤال خواهیم پرداخت که چرا داده‌کاوی مهم است؟ و لزوم انجام داده‌کاوی چیست؟

مزایای داده‌کاوی را می‌توان به شرح ذیل خلاصه نمود:

• فرض‌های ساده، مانند استقلال و یا وابستگی احتمالی همه پدیده‌ها را در نظر نمی‌گیرد.

• دینامیک است و محدود به توابع تعریف‌شده قبلی نیست؛ یعنی توابع به صورت دینامیک و دائماً از داده‌های خام موجود در مخزن داده ساخته می‌شوند. بنابراین، مدل محدود به تابع خاص نیست.

• آنالیزهای هم‌زمان بر روی پدیده‌های مختلف صورت می‌گیرد.

• به بومی‌سازی مدل نیازی نیست؛ زیرا مدل‌ها و توابع دائماً از بتن داده‌ها ساخته می‌شود و چون داده‌ها متعلق به سازمان است، بومی‌سازی معنا ندارد.

داده‌کاوی، همه چیز به همه چیز مرتبط است و آنهایی که به هم نزدیک‌ترند، مرتبط‌ترند.

شش عمل و وظیفه مهم را می‌توان برای داده‌کاوی به شرح ذیل برشمرد که سه مورد اوّل در گروه داده‌کاوی هدایت‌شده و دو مورد بعدی، در گروه داده‌کاوی غیرهدایت‌شده هستند:

ـ دسته‌بندی(۱۲)؛

ـ تخمین(۱۳)؛

ـ پیش‌بینی(۱۴)؛

ـ گروه‌بندی شباهت(۱۵)؛

ـ خوشه‌بندی(۱۶)؛

ـ توصیف و نمایه‌سازی(۱۷).

**متن‌کاوی، حوزه‌ای نو و میان‌رشته‌ای است که از رشته‌های بازیابی اطلاعات، داده‌کاوی، یادگیری ماشینی، آمار و زبان‌شناسی محاسباتی مشتق شده است**

داده‌کاوی، یکی از مؤلفه‌های مهم مدیریت، تحلیلی ارتباط با مشتری است و هدف آن، ارتباط مؤثر با مشتری است که به بازتولید روابط یادگیرنده می‌انجامد. تعاملات یک بانک با مشتریانش، داده‌های بسیاری را تولید می‌کند. این داده‌ها در وهله اوّل، از طریق سیستم‌های پردازش تعاملات مثل فایل اطلاعات مشتریان تولید می‌شود. سپس، داده‌ها جمع‌آوری، تصفیه خودکار و خلاصه می‌شوند تا در انبارداده‌های مشتری قرار گیرند (شهرابی، ۱۳۸۶).

### کاربردهای داده‌کاوی

رسالت اصلی داده‌کاوی، در دو طبقه کلی جای می‌گیرد: توصیف و پیش‌بینی. در سطح توصیف، هدف فهمیدن داده‌های زمان گذشته و حال است. از الگوهای توصیف برای جست‌وجوی گروهی از متغیرهای مشابه در افراد یا دسته‌هایی از گروه‌های جمعیت‌شناختی مشترک که ویژگی‌های خاصی از خود نشان می‌دهند، استفاده می‌شود. از پیش‌بینی نیز برای اظهارنظر درباره امور ناشناخته بر اساس امور شناخته‌شده استفاده می‌شود. از این ویژگی می‌توان برای پیش‌بینی آینده و یا اظهارنظر درباره حال استفاده کرد. در پیش‌بینی، دو نوع کارکرد وجود دارد: «رده‌بندی» که هدف از آن، قراردادن یک فقره در یک طبقه است و «تخمین» که هدف از آن، تولید مقادیر عددی برای یک متغیر ناشناخته است.

کاربردهایی که برای داده‌کاوی وجود دارند، بسیار گسترده‌اند و ما در این نوشتار، فقط امکان معرفی تعداد محدودی از آنها را داریم؛ به عنوان مثال‌های بیشتر، می‌توان به کاربردهای داده‌کاوی در زمینه‌های ذیل اشاره کرد:

- سیستم‌های مدیریتی؛ مثلاً مدیریت ارتباط با مشتریان؛
- نرم‌افزارهای امنیتی؛ مثلاً نرم‌افزاری مانیتورینگ شبکه و ویروس‌کش‌ها؛
- سیستم‌های بانکی؛ مثلاً تخصیص اعتبار به مشتریان و طبقه‌بندی آنها؛
- مالی و اقتصادی؛ مثلاً پیش‌بینی قیمت یک یا چند سهام یا شاخص؛
- برنامه‌ریزی و مکان‌یابی؛ مثلاً چینش داخلی فروشگاه‌های بزرگ و یا تخصیص امکانات شهری؛
- علوم پزشکی؛ مثلاً پیش‌بینی خطرات احتمالی ناشی از یک عمل جراحی خاص؛
- علوم اجتماعی و سیاسی؛ مثلاً پیش‌بینی یا تحلیل نتایج انتخابات.

**پزشکی:**

- تعیین نوع رفتار با بیماران و پیشگویی میزان موفقیت اعمال جراحی؛
- تعیین میزان موفقیت روش‌های درمانی در برخورد با بیماری‌های سخت؛
- تشخیص بیماری‌ها براساس انواع اطلاعات (تصاویر پزشکی، مشخصات بیمار احتمالی)؛

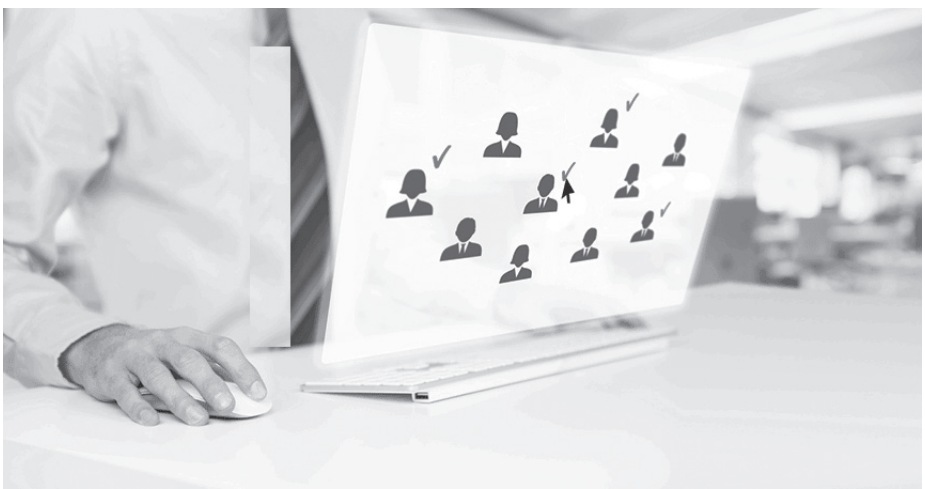
**متن‌کاوی، شبیه داده‌کاوی است؛ مگر اینکه ابزار داده‌کاوی برای کار با دادگان ساختاریافته از پایگاه دادگان طراحی شده است؛ اما متن‌کاوی ممکن است به مجموعه دادگان ساختارنیافته یا نیمه‌ساختاریافته مانند: ایمیل‌ها، اسناد متن کامل و مدارک اچ.تی.ام.ال نیز اعمال شود**

- تشخیص ناهنجاری‌هایی که توسط انسان به‌سختی قابل تشخیص خواهند بود.

**کتابخانه و داده‌کاوی**

در عصر کنونی، یکی از عناصر کلیدی و مهم سازمان‌ها برای حفظ بقا و چابکی سازمانی و همچنین مزیت رقابتی با دیگر سازمان‌ها، دانش است (کرمی، ۱۳۸۶). همان‌گونه‌که در بخش‌های قبلی عنوان شد، امروزه داده‌ها قلب تپنده هر سازمانی هستند؛ به این دلیل که بیشتر فعالیت‌های سازمان‌ها از طریق تعاملات در سیستم‌های عملیاتی شکل می‌گیرند و در نتیجه، نیاز است که این داده‌های ذخیره‌شده به‌خوبی پردازش شوند و اطلاعات حاصل از آن در اختیار کاربران قرار گیرد (شهرابی، ۱۳۸۶). سازمان‌ها در دنیای پُر رقابت امروزی، باید از لحاظ دانشی قوی باشند تا بتوانند به حیات و رقابت خود ادامه دهند. کتابخانه‌ها همانند بسیاری از سازمان‌ها با تغییرات سریع و پُرشتاب روبه‌رو هستند و در نتیجه، باید قدرت تحلیل شرایط و موقعیت‌های فعلی و آینده خود را داشته باشند تا بتوانند برای آینده خویش تصمیم‌گیری درست و به‌موقع داشته باشند و این زمانی محقق خواهد شد که کتابخانه‌ها از لحاظ دانشی، قوی و غنی باشند (کرمی، ۱۳۸۶).

برای تولید دانش، نیاز است که تراکنش بین داده‌ها شناسایی شده و با برقراری رابطه بین آنها، دانش کشف شده، در امور مختلف از آن بهره‌برداری نمود. استفاده از داده‌کاوی و متن‌کاوی، یکی از راهکارهای تولید دانش از پایگاه‌های اطلاعاتی است. داده‌کاوی، نه تنها در صدد یافتن اطلاعات یا پاسخگویی به سؤالاتی است که در ذهن کاربر وجود دارد، بلکه دانش عمیقی را که در دل داده‌ها نهفته است نیز کشف می‌کند. یکی از اهداف کتابخانه‌ها، رفع نیازهای کاربران است. داده‌کاوی می‌تواند در این امر نقش مهمی داشته باشد. با روش‌های تحلیل داده و کشف روابط بین آنها می‌توان کاربران را راضی‌تر نمود و در مواقعی باعث جذب بیشتر آنان گردید.



**متن‌کاوی**

در عصر کنونی، متون، یکی از ابزارهای مهم برای تبادل اطلاعات هستند و بیشتر اطلاعات رسمی نیز از همین طریق انتقال پیدا می‌کنند. بنابراین، نیاز است تا اطلاعات موجود در این متون، به گونه‌ای استخراج و تحلیل شوند. متن‌کاوی، یک حوزه جدید و در حال رشد است که سعی دارد اطلاعات معناداری را از متون زبان طبیعی استخراج کند (ویتن(۱۸)، ۲۰۰۰). متن‌کاوی که گاه به جای آن از واژه‌های «کاوش داده‌های متنی(۱۹)» و یا «کشف دانش در متن(۲۰)» استفاده می‌شود، برای نخستین‌بار در سال ۱۹۹۵ توسط فلدمن(۲۱) به کار برده شد (صالحی شهرودی و دیگران، ۱۳۹۲). متن‌کاوی،

به فرایندهایی از استخراج اطلاعات مطلوب و غیربدیهی، و نیز استخراج دانش از متون ساختارنیافته(۲۲) مرتبط می‌شود. متن‌کاوی، حوزه‌ای نو و میان‌رشته‌ای است که از رشته‌های بازیابی اطلاعات، داده‌کاوی، یادگیری ماشینی، آمار و زبان‌شناسی محاسباتی مشتق شده است. از آنجا که بسیاری از اطلاعات به شکل متن ذخیره شده‌اند، متن‌کاوی، ارزش اقتصادی بسیار بالایی در پی خواهد داشت. دانش، ممکن است از منابع گوناگون اطلاعاتی به دست آمده باشد، اما متون ساختارنیافته، بیشترین منابع دانش در دسترس را تشکیل می‌دهند. مسئله کشف دانش از متون، استخراج مفاهیم صریح و نیز غیرصریح و روابط معنایی میان مفاهیم، با استفاده از فنون پردازش زبان طبیعی تحقق می‌یابد. هدف استخراج دانش، به‌دست‌آوردن بصیرت‌هایی درباره دادگان متنی عظیم است. کشف دانش از متن، ریشه در پردازش زبان طبیعی دارد؛ اما روش‌هایی از آمار، یادگیری ماشینی، استدلال استخراج اطلاعات، مدیریت دانش و دیگر رشته‌های مرتبط برای فرایند کشف خود، از متن‌کاوی وام گرفته است.

متن‌کاوی، شبیه داده‌کاوی است؛ مگر اینکه ابزار داده‌کاوی برای کار با دادگان ساختاریافته از پایگاه دادگان طراحی شده است؛ اما متن‌کاوی ممکن است به مجموعه دادگان ساختارنیافته یا نیمه‌ساختاریافته مانند: ایمیل‌ها، اسناد متن کامل و مدارک اچ.تی.ام.ال نیز اعمال شود (ره‌آورد نور، ۱۳۹۳). تعاریف مختلفی از متن‌کاوی ارائه شده است؛ از جمله:

- متن‌کاوی به استخراج داده‌های متنی(۲۳) و کشف دانش از پایگاه‌های متن‌محور(۲۴) گفته می‌شود (فلدمن(۲۵)، ۱۹۹۵).
- به فرایند شناسایی اطلاعات جدید از مجموعه‌ای از متون گفته می‌شود (هیرست(۲۶)، ۱۹۹۹).
- متن‌کاوی به دنبال استخراج اطلاعات مفید از داده‌های متنی غیرساخت‌یافته از طریق تشخیص و نمایش الگوهاست؛ به عبارت دیگر، متن‌کاوی روشی برای استخراج دانش از متون است. متن‌کاوی، کشف اطلاعات جدید و از پیش ناشناخته، به وسیله استخراج خودکار اطلاعات از منابع مختلف نوشتاری است.
- کشف اطلاعات جدید و اطلاعاتی که از قبل ناشناخته بوده‌اند، توسط رایانه به کمک استخراج خودکار اطلاعات از منابع متنی غیرساخت‌یافتهٔ غالباً بزرگ.

در متن‌کاوی از همان تکنیک‌های داده‌کاوی استفاده می‌گردد و در کنار آن، به فناوری‌هایی مانند پردازش زبان طبیعی(۲۷) و یادگیری ماشین نیاز است تا به صورت خودکار، آمارهایی را جمع‌آوری نموده، ساختار و معنای مناسبی از متن استخراج گردد. بنابراین، متن‌کاوی= داده‌کاوی+ پردازش زبان طبیعی.

**تفاوت داده‌کاوی و متن‌کاوی**

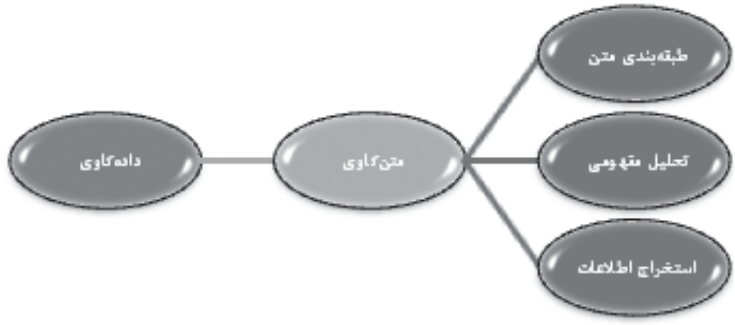
داده‌کاوی و متن‌کاوی، هر دو به دنبال کشف دانش از داده هستند. حال این سؤال مطرح می‌گردد که چه رابطه‌ای بین داده‌کاوی و متن‌کاوی وجود دارد؟

**استفاده از داده‌کاوی در مورد متن، شاخه‌ای دیگر را در علوم هوش مصنوعی به نام متن‌کاوی به وجود آورد که از جمله فعالیت‌های بسیار مهم در این زمینه، طبقه‌بندی (دسته‌بندی) متن است. در متن‌کاوی به دلیل اینکه روی متن کار می‌شود، باید از روش‌های پردازش زبان طبیعی استفاده کرد. در متن‌کاوی، کارهای پیش‌پردازش با استفاده ازروش‌های پردازش طبیعی و کارهای پردازش توسط داده‌کاوی انجام می‌شود**

داده‌کاوی، عبارت از کسب اطلاعات و دانش از یک مجموعه ساخت‌یافته. در داده‌کاوی، داده‌های ساخت‌یافته بررسی و تحلیل می‌شوند. داده‌های ساخت‌یافته، به داده‌هایی گفته می‌شود که به طور کاملاً مستقل از همدیگر، ولی یکسان از لحاظ ساختاری، در یک محل گردآوری شده‌اند. انواع پایگاه‌های اطلاعاتی را می‌توان نمونه‌ای از داده‌های ساخت‌یافته برشمرد؛ اما متن‌کاوی با متونی که عمدتاً غیرساخت‌یافته هستند، مانند صفحات وب، یادداشت، صورت‌حساب و ایمیل، و یا متون نیمه‌ساخت‌یافته، مانند صفحات اچ.تی.ام.ال(۲۸) و ایکس.ام.ال(۲۹) مواجه است. بنابراین، در متن‌کاوی، ابتدا باید متون را توسط

روش‌هایی ساختارمند نمود و سپس، از این روش‌ها برای استخراج اطلاعات و دانش از آنها استفاده کرد.

در نتیجه، متن، خود یک داده است و متن‌کاوی، یکی از زیرشاخه‌های داده‌کاوی است. استفاده از داده‌کاوی در مورد متن، شاخه‌ای دیگر را در علوم هوش مصنوعی به نام متن‌کاوی به وجود آورد که از جمله فعالیت‌های بسیار مهم در این زمینه، طبقه‌بندی (دسته‌بندی) متن است. در متن‌کاوی به دلیل اینکه روی متن کار می‌شود، باید از روش‌های پردازش زبان طبیعی استفاده کرد. در متن‌کاوی، کارهای پیش‌پردازش با استفاده ازروش‌های پردازش طبیعی و کارهای پردازش توسط داده‌کاوی انجام می‌شود (کروئزه(۳۰)، ماتئه(۳۱) و بودما(۳۲)، ۲۰۰۴).




**شکل٤: تفاوت داده‌کاوی و متن‌کاوی**

**تفاوت متن‌کاوی با بازیابی اطلاعات(۳۳)**

در بازیابی اطلاعات، برخلاف متن‌کاوی، هیچ اطلاعات جدیدی پیدا نمی‌شود و اطلاعات مورد نظر و مطلوب به‌ندرت با اطلاعات مشابه دیگری به طور هم‌زمان وجود دارند. متن‌کاوی، ترکیبی از فناوری‌های آماری، بازیابی اطلاعات، وب‌کاوی، داده‌کاوی و پردازش زبان طبیعی است. معمولاً در بازیابی اطلاعات، با توجه به نیاز مطرح‌شده از سوی کاربر، مرتبط‌ترین متون و مستندات و یا در واقع «کیسه کلمه» از میان دیگر مستندات یک مجموعه بیرون کشیده می‌شود. بازیابی اطلاعات، یافتن دانش نیست؛ بلکه تنها آن مستنداتی را که مرتبط‌تر به نیاز اطلاعاتی جست‌وجوگر تشخیص داده، به او تحویل می‌دهد. این روش، در واقع، هیچ دانش و حتی هیچ اطلاعاتی را به ارمغان نمی آورد.

متن‌کاوی، ربطی به جست‌وجوی کلمات کلیدی در وب ندارد. این عمل در حوزه بازیابی اطلاعات گنجانده می‌شود. به عبارتی 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 متن‌کاوی مجموعه‌ای از مستندات بررسی شده و اطلاعاتی که در هیچ‌یک از مستندات به صورت مجرد یا صریح وجود ندارد، استخراج می‌گردد.

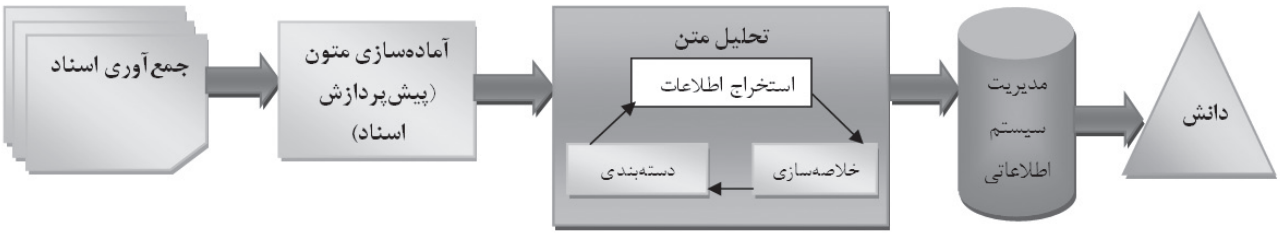
**مراحل متن‌کاوی**

متن‌کاوی، شامل سه مرحله است که عبارت‌اند از:

**۱. آماده‌سازی متن:** این مرحله، انتخاب، پاک‌سازی و پردازش مقدماتی متن را شامل می‌شود. در این مرحله، پایگاه‌ها یا منابعی که قرار است متن‌کاوی بر روی آنان انجام پذیرد، انتخاب می‌شوند که معمولاً این کار، با راهنمایی یک متخصص انسانی و یا نرم‌افزار مناسب انجام می‌شود. پردازش مقدماتی متن، از قبیل شناسایی جمله/ پاراگراف و برچسب‌گذاری نقش کلمه، در این مرحله انجام می‌شود.

**۲. پردازش متن:** این مرحله، شامل استفاده از الگوریتم داده‌کاوی به منظور پردازش داده‌های آماده‌سازی شده و فشرده‌سازی و انتقال آن، به منظور شناسایی قطعات پنهان اطلاعات است. در این فرایند، با استفاده از یک نظام پردازش زبان طبیعی، مشخصات استاندارد و گوناگون موجودیت‌ها (همچون: افراد، شرکت‌ها و سازمان‌ها) شناسایی می‌شود، رابطه مفهومی بین آنها مشخص می‌شود و حتی قالب‌های خاص مورد علاقه نیز معرفی می‌گردند. طبقه‌بندی شرکت‌کنندگان، تاریخ‌ها و نتایج، و جداول موجودیت‌ها و روابط استخراج‌شده، ویژگی‌های معناداری نظیر: درخت‌های تصمیم‌گیری،

**تعیین واژگان کلیدی و مهم متن، مهم‌ترین بخش فرایند متن‌کاوی است و دقت تمامی کاربردهای متن‌کاوی، تا حدود بسیاری به دقت در تعیین واژگان کلیدی بستگی خواهد داشت. برای تعیین واژگان کلیدی و مهم متن، الگوریتم‌های مختلفی بر اساس تکنیک‌های ریاضی ارائه شده است**




**شکل ۵: مراحل متن‌کاوی**

شبکه‌های خنثا، قوانین وابستگی یا الگوریتم‌های ژنتیک، برای الگوریتم‌ها و فنون استاندارد تهیه می‌کند.

**۳. تحلیل متن:** در این مرحله، برون‌داد مورد ارزیابی قرار می‌گیرد تا مشخص شود که آیا کشف دانش صورت پذیرفته است؟ و آیا دانش کشف‌شده اهمیت دارد یا خیر؟ با اجرای الگوریتم‌ها، داده/ متن استخراج‌شده به فنون مختلفی تحویل داده می‌شود که امکان استفاده مستقیم از اطلاعات استخراج‌شده را از طریق ابزار کشف پیوند یا مصورسازی فراهم می‌کنند.

این سه مرحله، باید به روشی اندیشمندانه صورت پذیرد؛ به‌طوری‌که به اهداف یک فرایند خاص متن‌کاوی، محدودیت‌های داده‌ها/ متن استخراج‌شده و نقاط قوت و ضعف الگوریتم مورد نظر، توجه کافی شود. شواهد نشان داده است که چنانچه این ملاحظات اعمال گردد، هم اطلاعات مرتبط و هم اطلاعات غیرمرتبط، کشف خواهد شد و در این صورت است که نتایج غیرمنتظره‌ای به وقوع خواهد پیوست و این، همان هدف متن‌کاوی، داده‌کاوی و همه انواع کشف دانش از داده‌هاست (خاصه، ۱۳۸۹).

### تکنیک‌های متن‌کاوی

این تکنیک‌ها عبارت‌اند از:

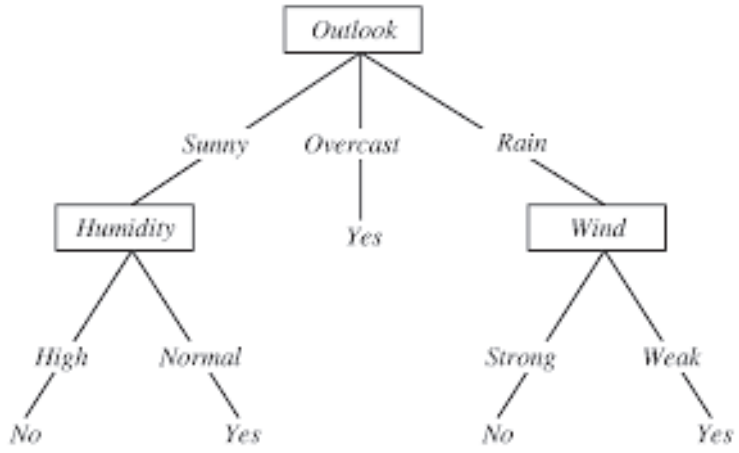
- **ساختارهای مفهومی.**
- **قوانین انجمنی(۳٤):** روابط و وابستگی‌های متقابل بین مجموعه بزرگی از اقلام داده‌ای را نشان می‌دهند. پیداکردن چنین قوانینی، می‌تواند در حوزه‌های مختلف مورد توجه بوده، کاربردهای متفاوتی داشته باشد؛ به عنوان مثال، کشف روابط انجمنی بین حجم عظیم تراکنش‌های کسب‌وکار، می‌تواند در تشخیص تقلب در حوزه پزشکی، و همچنین داده‌کاوی در مورد اطلاعات، روش به‌کارگیری وب توسط کاربران و شخصی‌سازی، مورد استفاده قرار گیرد (پارک(۳۵)، چن(۳۶) و یو(۳۷)، ۱۹۹۶) و یا در طراحی کاتالوگ، بازاریابی و دیگر مراحل فرایند تصمیم‌گیری کسب‌وکار، موثر باشد.
- **درخت‌های تصمیم‌گیری(۳۸):** درخت‌های تصمیم‌گیری، ابزار استانداردی در متن‌کاوی هستند. این الگوریتم‌ها، در متغیرها و نیز اندازه

**متن‌کاوی، ربطی به جست‌وجوی کلمات کلیدی در وب ندارد. این عمل در حوزه بازیابی اطلاعات گنجانده می‌شود؛ به عبارتی، بازیابی اطلاعات جست‌وجو، کاوش، طبقه‌بندی و فیلتر نمودن اطلاعاتی است که در حال حاضر شناخته شده‌اند و در متن قرار داده شده است؛ ولی در متن‌کاوی مجموعه‌ای از مستندات بررسی شده و اطلاعاتی که در هیچیک از مستندات به صورت مجرد یا صریح وجود ندارد، استخراج می‌گردد**

مجموعه آموزش، سریع و مقیاس‌پذیر هستند. یکی از مشکلات درخت‌های تصمیم‌گیری برای متن‌کاوی، این است که تنها به تعداد کمی از ترم‌ها وابسته است.

◄ **روش‌های استنتاج قوانین.**

◄ **تکنیک‌های بازیابی اطلاعات.**




**شکل٦: نمونه‌ای از درخت تصمیم‌گیری**

**استخراج مفهوم(۳۹)(معنا) از متن**

متن‌کاوی، وابسته به زبان و زبان‌شناسی است و سعی در کشف مفاهیم و معانی موجود در متن به جای شمارش واژه‌های پُرتکرار در متن دارد. استخراج مفاهیم، هم رایانه و هم انسان را که در یک همکاری دوسویه هستند، شامل می‌گردد. قبل از پرداختن به استخراج 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 بر اساس هدف و نیاز خود، تعاریف مختلفی از معنا را ارائه می‌دهند؛ اما همه آنها به نوعی وام‌دار و وابسته به زبان‌شناسی هستند. در زبان‌شناسی، هر واژه دارای یک معنای ثابت بوده، 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 (رشیدیان، ۱۳۸۳). 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 است و کاربر قصد دارد اطلاعاتی را درباره آن جست‌وجو نماید (پارامسواران(۴۰) و دیگران، ۲۰۱۰).

* **استخراج معنا از متن:** فرض کنید که یک فرد دارای یک کیسه حاوی کلمات نامرتبی است که متن یک جمله یا یک پاراگرف را تشکیل می‌دهد. از در کنار یکدیگر قرار دادن این کلمات، معنا دریافت می‌گردد. 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 نمایند؛ اما رایانه‌ها قادر به این امر نیستند و باید از قبل یکسری از اطلاعات برای آنها تعریف گردد تا بتوانند بخشی از این کار را انجام دهند.

استخراج مفهوم، یکی از مهم‌ترین ابزارهای طبقه‌بندی متن بر اساس مفاهیم است (یونتائو(۴۱) و دیگران، ۲۰۰۳). استخراج معنا از یک متن، علاوه بر اینکه جست‌وجو را گسترش می‌دهد، تقویت‌کننده ارائه مدارک مرتبط‌تر با اصطلاح مورد جست‌وجو نیز هست و دارای فواید دیگری است که شامل موارد ذیل است:

• ارائه مفاهیم مرتبط با اصطلاح مورد جست‌وجو. این فایده، کاربر را با مفاهیم دیگری آشنا کرده، در نتیجه، دانش جدیدی را در اختیار وی قرار می‌دهد.

• ایجاد روابط بین مفاهیم و گسترش دامنه جست‌وجو؛

• توسعه دانش زمینه‌ای؛

• تولید تفاسیر متون؛

• تعریف روابط بین متون؛

**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 نمایند؛ اما رایانه‌ها قادر به این امر نیستند و باید قبلاً یکسری از اطلاعات برای آنها تعریف گردد تا بتوانند بخشی از این کار را انجام دهند**

• طبقه‌بندی و دسته‌بندی متون (پارامسواران(۴۲) و دیگران، ۲۰۱۰).

**مشکلات استخراج معنا از متن**

۱. مفاهیم می‌تواند در قالب یک کلمه و 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 مثال، کلمه «شبکه»، یک مفهوم، و «شبکه معنایی»، مفهوم خاص‌تری دارد. بنابراین، یکی از چالش‌های استخراج مفاهیم، توجه به این مورد است.

۲. در یک زمینه موضوعی خاص نیز مفاهیم یک کلمه و چند کلمه‌ای به‌کار می‌رود که برای گزینش آنها باید به میزان تکرارشان در متن توجه نمود.

**چگونگی استخراج معنا از متن**

هر متن، دارای یکسری از کلمات و واژه‌هاست که برخی از آنها، فاقد بار اطلاعاتی هستند و برخی دیگر، مفاهیم اصلی را بازگو می‌کنند. بسیاری از کلمات همچون حروف اضافه و حروف ربط که در متن بسیار تکرار می‌شوند، قبل از استخراج معنا و مفاهیم از متن، در فهرست‌های بازدارنده(۴۳) نگهداری می‌شوند و هنگام استخراج مفاهیم از متن، حذف می‌گردند. یکی دیگر از گام‌ها قبل از استخراج معنا، پیش‌پردازش متن است؛ برای مثال، در زبان انگلیسی، شناسایی کلمات هم‌ریشه؛ مانند developing و developed که از ریشه develop است. دلیل پیش‌پردازش متن، آماده‌سازی آن برای انجام استخراج مفاهیم اصلی موجود در متن است. بین کلمات موجود در متن، روابط مختلفی نیز حاکم است؛ مثلاً هم‌معنا، مترادف و شبه مترادف. در استخراج معنا، این روابط شناسایی شده، همه به عنوان یک کلمه در نظر گرفته می‌شوند.

در آماده‌سازی داده‌های متنی، چگونگی ذخیره داده‌های غیرساخت‌یافته، نمایه‌سازی آنها و تبدیل فرمت آنها به شکل‌های قابل استفاده در نرم‌افزارها، مورد بررسی قرار می‌گیرد. نادیده گرفتن واژگان بدون محتوای باارزش اطلاعاتی، همانند: از، به، با، تا و... که معمولاً بیشترین میزان تکرار در متن را به خود اختصاص می‌دهند، در مرحله دوم مورد بررسی قرار می‌گیرد. از آنجا که در متن‌کاوی به دنبال واژگان ارزشمند اطلاعاتی هستیم، این دسته از واژگان را کنار خواهیم گذاشت. ریشه‌یابی واژگان، در مرحله سوم مورد بررسی قرار می‌گیرد. واژگانی مانند: «دانش»، «دانشمند» و «دانایی»، واژگان هم‌ریشه به حساب می‌آیند. در بررسی یک متن توسط فرایند متن‌کاوی، به تعیین ریشه و واژگانی هم‌ریشه پرداخته می‌شود تا در بررسی آماری واژگان متن، این دسته از واژگان در کنار هم مورد بررسی قرار گیرد.

تعیین واژگان کلیدی و مهم متن، مهم‌ترین بخش فرایند متن‌کاوی است و دقت تمامی کاربردهای متن‌کاوی، تا حدود بسیاری به دقت در تعیین واژگان کلیدی بستگی خواهد داشت. برای تعیین واژگان کلیدی و مهم متن، الگوریتم‌های مختلفی بر اساس تکنیک‌های ریاضی ارائه شده است. روش تکرار واژگان، معمولی شده و روش وزن‌دهی بر اساس جست‌وجوی کاربر و سایر الگوریتم‌های مشتق‌شده

**شبکه‌های واژگانی، ابزارهای ارزشمندی در پردازش زبان طبیعی هستند که برای نخستین‌بار بر مبنای یافته‌های روان‌شناسی زبان ساخته شدند. این شبکه‌ها، بر اساس روابط معنایی میان واژه‌ها شکل گرفته‌اند و تلاشی در جهت بازنمایی آنچه در ذهن انسان‌ها از واژه‌ها و روابط آنها وجود دارد، می‌باشند**

از این الگوریتم‌ها در تعیین واژگان کلیدی و مهم به‌کار گرفته می‌شوند. وجه اشتراک تمامی این الگوریتم‌ها، آن است که تعیین واژگان کلیدی یک متن، بر اساس آنالیز یک مجموعه متون مرتبط با هم تعیین می‌شود. کاربردهای مختلفی را می‌توان از فرایند، «کا ـ مینز» متن‌کاوی انتظار داشت. بر اساس واژگان کلیدی می‌توان با الگوریتم‌های خوشه‌بندی و یا سایر الگوریتم‌ها به خوشه‌بندی متون پرداخت. با استفاده از خوشه‌بندی متون «سلسله‌مراتبی» می‌توان متن‌های مشابه را شناسایی و دسته‌بندی کرد. با تعیین واژگان کلیدی و استفاده از تکنیک‌های مختلف وزن‌دهی به جملات و واژگان متن، می‌توان یک چکیده از متن تهیه و ارائه کرد (زعفریان، ۱۳۸۵).

**رویکردهای معمول استخراج مفاهیم از متن**

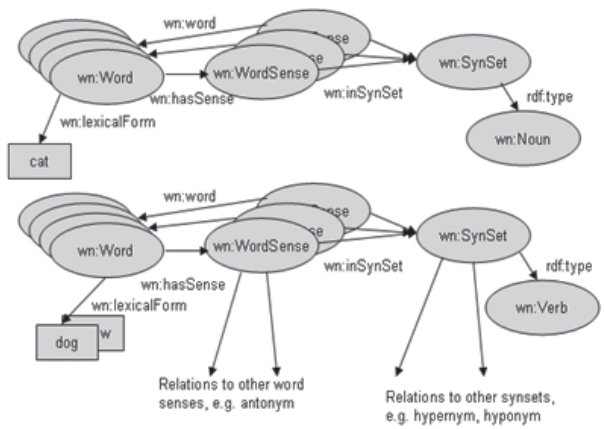
از روش‌هایی که در سال‌های گذشته و اکنون در زمینه دریافت معنا از متن در رایانه به‌کار می‌رود، روش‌ها و رویکردهای مربوط به زبان‌شناسی و دانش‌محور است. این رویکردها، به ساختار دانشی متن و هستی‌شناسی‌ها مربوط هستند که

بیشتر روی زبان و ساختار آن تأکید می‌کنند که در ادامه به برخی از آنان پرداخته می‌شود:

* **شبکه‌های واژگانی:** شبکه‌های واژگانی، ابزارهای ارزشمندی در پردازش زبان طبیعی هستند که برای نخستین‌بار بر مبنای یافته‌های روان‌شناسی زبان ساخته شدند. این شبکه‌ها، بر اساس روابط معنایی میان واژه‌ها شکل گرفته‌اند و تلاشی در جهت بازنمایی آنچه در ذهن انسان‌ها از واژه‌ها و روابط آنها وجود دارد، می‌باشند و بنابراین، تلاش بر این است که حداکثر واژه‌های موجود در یک زبان را به صورت شبکه‌ای از روابط در خود بگنجانند (حسابی، ۱۳۹۴)؛ برای مثال، وردنت، یک پایگاه داده لغوی بزرگ از لغات انگلیسی است. این بانک اطلاعاتی، اسم‌ها، فعل‌ها، صفت‌ها و قیدها را به مجموعه ای از لغات مترادف دسته‌بندی می‌نماید که هر دسته، یک مفهوم مجزا را بیان می‌کند. مجموعه مترادف‌ها، با استفاده از روابط معنایی ـ مفهومی و ارتباطات لغوی به یکدیگر پیوند داده شده‌اند. شبکه به‌دست‌آمده که شبکه‌ای است از لغات و مفاهیم مرتبط از لحاظ معنایی، می‌تواند توسط مرورگرها پیمایش شود. به علاوه، وردنت به صورت رایگان و برای عموم در دسترس و قابل بارگذاری است. ساختار وردنت، از آن یک ابزار مفید جهت زبان–شناسی محاسباتی و پردازش زبان طبیعی به وجود آورده است. وردنت، مشابه یک لغت‌نامه است که لغات را بر اساس معانی آنها دسته‌بندی می‌کند؛ هرچند تفاوت‌های مهمی بین وردنت و دیگر لغت‌نامه‌ها وجود دارد:

اوّل اینکه وردنت، تنها شکل کلمات ـ رشته‌هایی از حروف ـ را پیوند نمی‌دهد؛ بلکه مفاهیم لغات را نیز مرتبط می‌سازد. در نتیجه، 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 نمی‌نماید.

* **هستی‌شناسی(٤٤):** هستی‌شناسی، الگویی انتزاعی از جهان واقع است که




**شکل ۷:**

**Source: www.w3.org/.../WNET/wordnet-sw-20040713.html**

**روابط معنایی در حوزه متن می‌توانند در سطوح پایین و بین واژگان و یا در سطوح بالاتر بین عبارات، جمله‌ها، پاراگراف‌ها و حتیٍ بالاتر از بخش‌های یک متن، مثلاً بین دو سند یا مجموعه‌ای از اسناد رخ دهد. سطوح ذکرشده، دارای یک محدوده تعریف ساختاری می‌باشند**

مفاهیم و روابط میان آن را در قلمروی مورد بحث نمایش می‌دهد. هستی‌شناسی‌ها، پایگاه دانش مفهومی(۴۵) هستند (شمس فرد و عبدالله‌زاده، ۱۳۸۱). هستی‌شناسی‌ها به منزله ابزار بازنمون دانش در نظام‌های ذخیره و بازیابی، استفاده می‌شوند و آن را مجموعه‌ای از مفاهیم، خصیصه‌ها و روابط میان آن مفاهیم تعریف کرده‌اند. این تعریف در حوزه الگوسازی مفهومی، چندان جدید نیست. مدل‌های موجودیت ـ رابطه از دهه ۱۹۷۰ در پایگاه‌های اطلاعاتی استفاده می‌شود و در الگوهای گسترش‌یافته آن نیز چنین الگویی از مفاهیم، خصیصه‌ها و روابط قابل شناسایی است؛ اما دلیل این همه استقبال از هستی‌شناسی‌ها در این نکته نهفته است که هستی‌شناسی‌ها برخلاف الگوهای مفهومی پیش‌گفته، استنتاج هوشمند را ممکن می‌سازند (شریف، ۱۳۸۸).

* **رویکرد آماری:** این رویکرد، به میزان تکرار یک کلمه در متن اشاره دارد.

## استخراج روابط معنایی در سطح گفتمان از متن

عمده کارها در این حوزه، به دو دسته «استخراج مفاهیم» و «استخراج روابط» تقسیم می‌شوند. بسیاری از کاربردها در استخراج اطلاعات،

**شکل ۸: نمونه‌ای از هستی‌شناسی**

اکتساب دانش، فهم زبان طبیعی و بازیابی اطلاعات، نیازمند درک در سطح مفهوم، معنا، رابطه و روابط معنایی است. هرکدام از موجودیت‌های ذکرشده، دارای ویژگی‌ها و تعاریف مشخص می‌باشند. رابطه معنایی، به روابط بین مفاهیم موجود در ذهن که به آن مفاهیم ادراکی گفته می‌شود یا به روابط بین کلمات که روابط لغوی نامیده می‌شوند، اشاره دارد. روابط معنایی در حوزه متن می‌توانند در سطوح پایین و بین واژگان و یا در سطوح بالاتر بین عبارات، جمله‌ها، پاراگراف‌ها و حتی بالاتر از بخش‌های یک متن، مثلاً بین دو سند یا مجموعه‌ای از اسناد رخ دهد. سطوح ذکرشده، دارای یک محدوده تعریف ساختاری می‌باشند.

اما سطح دیگری در حوزه زبان‌شناسی، سطح گفتمان نام دارد که محدوده تعریف آن، نه ساختار، بلکه معناست. در این سطح، مقیاس، نه یک جمله یا پاراگراف یا سند، بلکه واحد معنایی است و ممکن است بین چند جمله، چند پاراگراف و یا تمام متن باشد. نکته قابل توجه، وقوع این روابط معنایی در سطح گفتمان، به صورت صریح و یا ضمنی در متن است و اهمیت آن، به دلیل انتشار روزافزون اطلاعات در سطح وب است که غالباً در قالب متن هستند و اکتساب دانش و استخراج اطلاعات مورد نیاز از بین این حجم انبوه داده، مستلزم درک مفاهیم و روابط معنایی حاکم بین آنها، به‌ویژه در حوزه‌های زبان‌شناسی، روان‌شناسی، علوم‌شناختی و زبان‌شناسی محاسباتی است. بهره‌گیری از روش‌های پردازش زبان طبیعی در کنار محاسبات آماری و روش تجزیه مقدار تکینه و روش استخراج و شناسایی الگو بدون ناظر، از مشخصه‌های روش پیشنهادی است. هدف، کمک به درک و فهم دقیق و عمیق‌تر از متن و پیاده‌سازی در کاربردهای بسیار در حوزه‌های مرتبط با زبان‌شناسی، روان‌شناسی، علوم‌شناختی و به‌ویژه زبان‌شناسی محاسباتی است. تحقیقات موجود نشان می‌دهد که روش‌های ارائه‌شده هوش مصنوعی و یادگیری ماشین در این حوزه، نیازمند سطح بالایی از مهندسی دانش یا پایگاه داده‌های بزرگ برای یادگیری و آزمایش هستند (کاهانی، ۹۲).

**یادگیری از متن‌ها**

این نوع یادگیری، شامل استخراج هستی‌شناسی با استفاده از تکنیک‌های پردازش زبان طبیعی از متن است. دستاوردهای شناخته‌شده در این گروه، به صورت ذیل می‌باشد:

* **استخراج مبتنی بر الگو:** یک رابطه وقتی شناخته می‌شود که یکسری از کلمات در متن به یک الگو نگاشت شوند؛ برای مثال، یک الگو می‌تواند ایجاد شود، اگر یک دنباله‌ای از n نام داریم. سپس، n–۱ از این نام‌ها hyponyms برای nامین نام می‌باشند.

* **قوانین انجمنی:** این قوانین در ابتدا در زمینه پایگاه داده به شرح ذیل تعریف شد:

با توجه به مجموعه‌ای از تراکنش‌ها که هر تراکنش مجموعه‌ای از آیتم‌هاست. یک قانون انجمنی، یک عبارت از شکل X دلالت می‌کند به Y است که در آن، X و Y مجموعه از آیتم‌ها هستند. معنای مستقیم چنین قانونی، این است تراکنش‌هایی از پایگاه داده که شامل X است، گرایش به Y دارند. قوانین انجمنی، در فرآیند داده‌کاوی به منظورکشف اطلاعات ذخیره‌شده در پایگاه داده مورد استفاده قرار می‌گیرد.

* **خوشه‌بندی مفهومی:** مفاهیم با توجه به فاصله معنایی بین یکدیگر برای ساختن سلسله‌مراتب، گروه‌بندی می‌شوند. فرمول برای محاسبه فاصله معنایی بین دو مفهوم، ممکن است به عوامل مختلفی بستگی داشته باشد که در این روش‌ها ارئه می‌شود.

* **یادگیری مفهوم:** یک طبقه‌بندی داده‌شده به صورت تدریجی به‌روز می‌شود؛ به‌طوری‌که مفاهیم جدید، از متن دنیای واقعی استخراج می‌شوند.

* **یادگیری هستی‌شناسی از فرهنگ لغت:** در ساخت هستی‌شناسی، از یک دیکشنری که قابل خواندن توسط ماشین است، برای استخراج مفاهیم مرتبط و روابط بین آنها استفاده می‌شود که هدف، ساخت یک هستی‌شناسی خلاصه‌شده است.

* **یادگیری از یک پایگاه دانش:** هدف این کار، یادگیری هستی‌شناسی با استفاده از منبع موجود پایگاه دانش است.

* **یادگیری از داده‌های نیمه‌ساخت‌یافته:** یعنی استخراج هستی‌شناسی از منابعی که ساختار از پیش تعریف‌شده دارند؛ مانند شمای XML.

* **یادگیری از شمای رابطه‌ای:** هدف این یادگیری، استخراج مفاهیم مرتبط هستی‌شناسی و روابط از دانش پایگاه داده است.

**پی‌نوشت‌ها:**

1. data Mining.
2. Davenport.
3. Knowledge Discovery.
4. Fyyad.
5. Hotho.
6. knowledge discovery in databases (KDD).
7. Berson.
8. Gyorodi.
9. Berry.
10. Data Cleaning.
11. pattern evaluation.
12. Classification.
13. Estimation.
14. Prediction.
15. Affinity Grouping.
16. Clustering.
17. Profiling.
18. Witten.
19. text data mining.
20. Knowledge Discovery in Text.
21. Feldman.
22. Unstructured.
23. Text Data Mining (TDM).
24. Textual Database.
25. Feldman.
26. Hearst.
27. natural language processing (NLP).
28. HTML.
29. XML.
30. Kroeze.
31. Matthee.
32. Bothma.
33. Information Retrieval.
34. association rule.
35. Park.
36. Chen.
37. Yu.
38. Decision tree.
39. Concept extraction.
40. Parameswaran.
41. Yuntao.
42. Parameswaran.
43. Stop list.
44. Ontology.
45. conceptual knowledge.

**منابع فارسی:**


۱. باقری، شیرین و سنجر سلاجقه. «از مدیریت داده تا مدیریت دانش». عصر مدیریت. ۱۴. (۱۳۸۹): ۷۶ ـ ۸۱.

۲. تاج‌الدینی، اورانوس، علی سادات موسوی و ساره دلیری. «داده‌کاوی: تعمیق نگاه برای کشف دانش در پیوند با دنیای الکترونیک و کاربرد آن در کتابداری و اطلاع‌رسانی». ماهنامه اطلاع‌یابی و اطلاع‌رسانی. ۲۱. (۱۳۸۸): ۲۷ ـ ۳۳.

۳. حسابی، اکبر. «مقایسه روابط معنایی درون‌زبانی اسامی در فارس‌نت، یورونت و وردنت پرینستون». دوماهنامه جستارهای زبانی. ۴. (۱۳۹۵): ۱۴۹ ـ ۱۷۳.

۴. خاصه، علی‌اکبر. «داده‌کاوی، متن‌کاوی و وب‌کاوی؛ تعاریف و هدف‌ها». مجله الکترونیکی ارتباط علمی. ۲. (۱۳۸۹): ۱۶.

۵. رشیدیان، عبدالکریم. «دریدا و نظریه معنا در هوسرل». پژوهشنامه علوم انسانی دانشگاه شهید بهشتی. ۳۹ و ۴۰. (۱۳۸۲): ۷۳ ـ ۸۶.

۶. «کاربرد متن‌کاوی در سازمان‌دهی دانش». ره‌آورد نور. ۴۸. (۱۳۹۳): ۲۳.

۷. زعفریان، ر. ۱۳۸۵. «روش‌هایی برای متن‌کاوی متون فارسی به همراه مطالعه موردی در مهندسی صنایع». رساله دکترا. تهران: دانشگاه صنعتی شریف.

۸. شهرابی، جمال. ۱۳۸۶. **داده‌کاوی**. تهران: مؤسسه پژوهشی داده‌پردازان گیتا و جهاد دانشگاهی واحد صنعتی امیرکبیر.

۹. شریف، عاطفه. «مهندسی خودکار هستی‌شناسی؛ امکان‌سنجی استخراج روابط معنایی از متون فارسی و تعیین میزان پیدایی آنها». فصلنامه کتابداری و اطلاع‌رسانی. ۴۶. (۱۳۸۸): ۲۴۳ ـ ۲۶۳.

۱۰. شمس فرد، مهرنوش و احمد عبدالله‌زاده. «استخراج دانش مفهومی از متن با استفاده از الگوهای زبانی و معنایی». تازه‌های علوم شناختی. ۱۳. (۱۳۸۱): ۴۸ ـ ۶۶.

۱۱. صالحی شهرودی، محمدحسین، بهروز مینایی و امیررضا اشرفی. «متن‌کاوی موضوعی رایانه‌ای قرآن کریم برای کشف ارتباطات معنایی میان آیات، بر مبنای تفسیر المیزان». قرآن شناخت. ۱۲. (۱۳۹۲): ۱۱۷ ـ ۱۵۲.

۱۲. کاهانی، محسن. ۱۳۹۲. «استخراج روابط معنایی در سطح گفتمان از متن». رساله دکترا. مشهد: دانشگاه فردوسی.

۱۳. کرمی، مهتاب. «کاربرد ابزارهای تحلیلگر داده‌کاوی و متن‌کاوی در چابکی سازمان‌های مراقبت بهداشتی و درمانی». فصلنامه علمی ـ پژوهشی مدیریت سلامت. ۳۰. (۱۳۸۶): ۱۵ ـ ۲۰.

## منابع انگلیسی:

1. Berry, Michael and Linoff, Gordon (1997) “Data Mining Techniques: For Marketing, Sales, and Customer Support” New York: John Wiley and Sons.

2. Berson Alex, Smith Stephen, and Thearling Kurt “Building Data Mining Applications for CRM”, 2004, Tata McGraw-Hill, New York.

3. Davenport, T.H. & Prusak, L .(1998). “ Working Knowledge: How organizations manage what they know “ Harvard business school press, boston.

4. Fayyad, U. (1996). Data mining and knowledge discovery: Making sense out of data. IFEE Expert, 66(5).

5. Feldman, R.(1995). knowledge discovery in texts. In Proc. of the First Int. Conf. on Knowledge Discovery (KDD).

6. Gyorodi, Robert S.(2001). A Comparative Study of Iterative Algorithm in Association

Kroeze, J.H., Matthee, M.C.; Bothma, T.J.D.(2004). Differentiating between data-mining and text-mining terminology.south Africa journal of information management, 6(4).

7. Hearst, M. (1999). Untangling text data mining. Proceedings of ACL’99: the 37th Annual Meeting of the Association for Computational Linguistics.

8. Hotho A, Nurnberger A, Paab G. Abrief survey of text mining. 2005. Available from:

http:// www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05 TextMiningæ.

9. Park, J.S.; Chen, M.S.; Yu, P.S. (1996).Data mining for Path Traversal Patterns in a Web Environment. Proceedings of the 16th International Conference on Distributed Computing

Parameswaran, Aditya; Garcia-molina, Hector; Rajaraman,Anand(2010). Towards the web of concepts: extracting concepts from large datasets. Proceedings of the VLDB Endowment, 3(1-2).

10. Robert S.Gyorodi, "A Comparative Study of Iterative Algorithm in Association Rules Mining ", Studies in Information and control, Vol.12, No.3.

11. Witten, I.H. and Frank, E. (2000) Data mining: Practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, San Francisco, CA.

12. Yuntao, Zhang; Ling, Gong; Yongcheng, Wang; Zhonghang, Yin(2003). An Effective Concept Extraction Method for Improving Text Classification Performance. Geo-spatial Information Science, 6(4).■