استخراج دانش (داده کاوی)

در دو دهه گذشته توانایی‌های فنی بشر برای تولید و جمع‌آوری داده‌ها به سرعت افزایش یافته‌است. عواملی نظیر استفاده گسترده از بارکد برای تولیدات تجاری، به خدمت‌گرفتن کامپیوتر در کسب و کار، علوم، خدمات دولتی و پیشرفت در وسائل جمع‌آوری داده از اسکن‌کردن متون و تصاویر تا سیستم‌های سنجش از دور ماهواره‌ای، در این تغییرات نقش مهمی داشته‌اند (تولید حجم انبوهی از داده‌ها). بنابراین امروز در روزگاری زندگی‌می‌کنیم که عصر انفجار اطلاعات نامیده‌می‌شود. همه روزه با حجم وسیعی از داده‌ها و اطلاعات پیرامون خود مواجه هستیم. اما حلقه‌ی مفقوده‌ی عصر ما دانش است. رشد روز افزون داده‌ها در شاخه‌های مختلف صنعت و علوم باعث شده است تا از کامپیوتر و علوم مربوط به آن جهت پردازش این حجم بالا از داده‌ها استفاده‌شود. هدف از پردازش داده‌ها، استخراج دانش از آنها به گونه‌ای است که بتوان در کاربردهای دیگر از آنها استفاده نمود.[1]

داده کاوی: علمی است که حجم عظیمی از داده‌ها را مورد پردازش عمیق قرارمی‌دهد تا نظم‌هایی را که در عمق داده‌ها وجود دارد- همچون طلا در یک معدن طلا –به صورت دانشی با ارزش کشف و جهت استفاده عرضه می‌کند.

انگیزه ها و دلایل اصلی ظهور داده‌کاوی

- رشد روز افزون داده‌ها (انفجار اطلاعات)
- معمولا دانشی که در داده‌ها وجود دارد، خیلی بدیهی و روشن نیست.
- کشف اطلاعات با ارزش از داده‌ها توسط تحلیل‌گران انسانی ممکن است هفته‌ها یا ماه‌ها زمان نیاز داشته باشد.

داده‌های مورد استفاده در داده کاوی عمدتا در یکی از دو نوع زیر خواهد بود 1- داده‌های تجاری 2- داده‌های علمی

سه منبع اصلی برای داده‌های تجاری قابل تصور است:

1) داده‌های وب یا داده‌های تجارت الکترونیک
2) خرید و فروش‌های موجود در فروشگاه‌ها
3) تراکنش‌های بانکی و کارت های بانکی

نکته: انگیزه اصلی داده‌کاوی در داده‌های تجاری، انگیزه مالی است که باعث‌می‌شود شرکت‌ها برای رقابت بهتر از کاوش داده‌ها بهره بگیرند.

به طور کلی برای جمع‌آوری داده‌های علمی چهارمنبع وجود دارد که عبارتند از:

1) تصاویر ارسالی از طریق ماهواره‌ها
2) تصاویر ارسالی از طریق تلسکوپ‌ها
3) داده‌های دنباله ژنی
4) داده‌های حاصل از شبیه‌سازی علمی

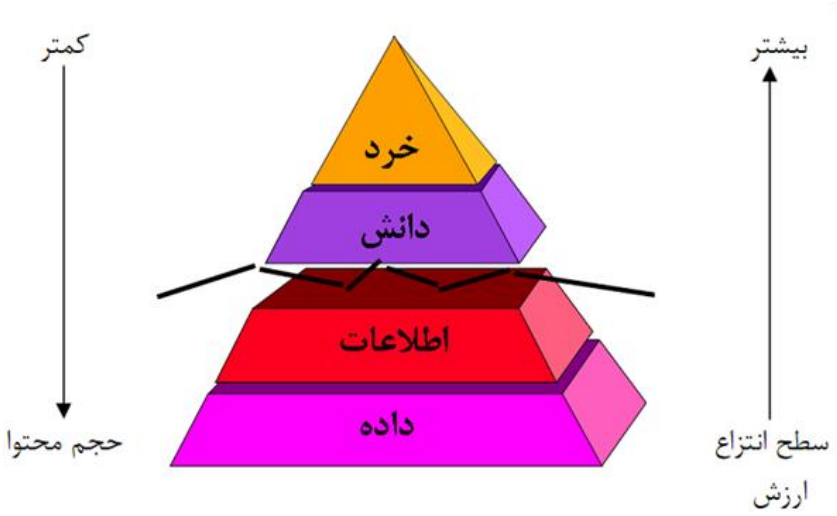
نکته: انگیزه اصلی در پردازش داده‌های علمی، گسترش مرزهای دانش بشری در یک حوزه خاص علمی است.

[1] - بخشی از مطالب سری نهم از وبلاگ مصطفی سبزه کار و بخشی هم از کتاب داده کاوی کاربردی- نویسندگان: صنیعی، محمودی و طاهرپرور- برگرفته‌شده‌است.

شرایط لازم برای داده‌کاوی 1- وجود داده 2- صحت داده 3- کافی بودن ویژگی‌هاست. برای فهم و درک بهتر داده‌کاوی لازم است تعاریف داده، اطلاعات و دانش، مرور و یادآوری شود.

- **داده:** هر گونه سمبل، عدد، رقم، کاراکتر، رشته و یا سیگنالی که معنای خاصی نداشته باشد(فاقد معنای خاصی باشد)
- **اطلاعات**: هر گونه سمبل، عدد، رقم، کاراکتر، رشته و یا سیگنالی که معنای خاصی داشته باشد(دارای معنای خاصی باشد)
- **دانش:** وجود رابطه بین دو عنصر اطلاعاتی بیان‌می‌کند.

در شکل 9-1 که هرم دانش نامیده‌می‌شود سلسله مراتب دانش، نشان‌داده‌شده‌است. همانطورکه از شکل9-1 پیداست برای رسیدن به دانش موجود در داده‌ها، راه سختی در پیش داریم.



شکل9-1: هرم دانش(یادآوری)

نکته: با افزایش ارزش معنایی مفاهیم، حجم آنها کاهش می‌یابد که کاملا طبیعی است. به عبارت دیگر حجم بالایی از داده‌ها را می‌توان تنها با چند قانون توصیف و تبیین‌کرد که از اهداف اصلی داده‌کاوی است.

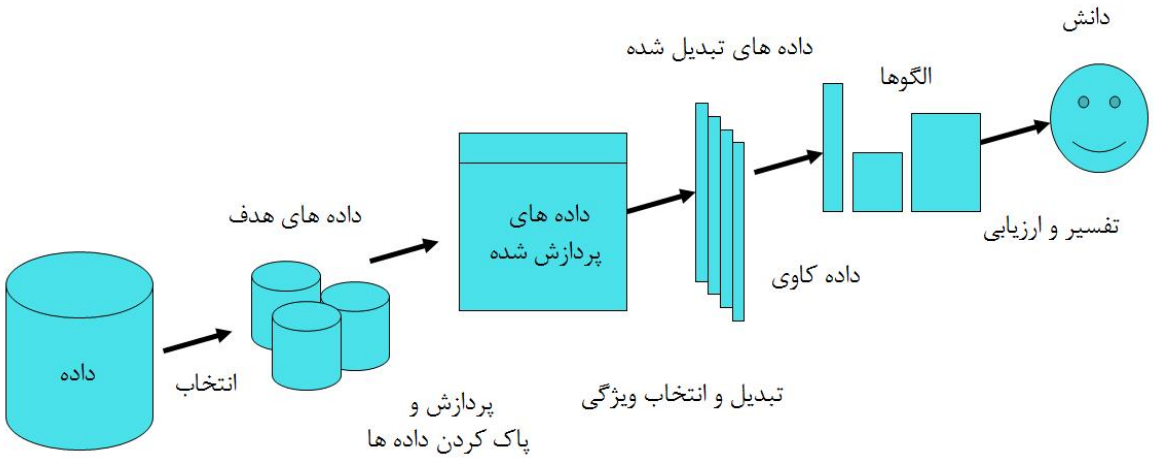
هدف اصلی در داده‌کاوی، کشف دانش است. دانش نظم موجود داده‌هاست. پس از کشف دانش ممکن است یکی از دو حالت زیر اتفاق بیفتد.

1- افراد خبره، آگاه به دانش استخراج شده باشند (دانش به صورت قانون تلقی‌می‌شود)
2- دانش کشف‌شده دانش جدید باشد. در این حالت، دانش جدید بررسی‌شده و در صورت منطقی بودن، به فرضیه تبدیل‌می‌شود. با آزمایش‌ها و بررسی‌های عمیق‌تر و بیشتر، درست یا نادرست بودن فرضیه اثبات‌می‌شود. اگر فرضیه اثبات‌شد، به قانون تبدیل‌خواهد شد.

می‌توان گفت که برای فرایند استخراج دانش، دو رویکرد اصلی قابل تصور است:

- این فرایند به صورت مهندسی انجام‌شود (یعنی توسط متدولوژی‌های مهندس دانش انجام شود) که معمولا اکتساب دانش یا دریافت دانش نامیده‌می‌شود. (فرم‌های خاصی پر شود، یا نمودارهایی ترسیم شود و ...) اکتساب دانش برای هر مساله‌ای امکان پذیر نیست.
- راه حل دیگر داده‌کاوی است که به صورت خودکار دانش، استخراج می‌شود که به آن اکتشاف دانش هم گفته‌می‌شود.

اکتشاف دانش (Knowledge Discovery = KD) عبارتست از پروسه‌ای جهت استخراج اطلاعات مهم و اساسی، ضمنی، قبلا ناشناخته و سودمند از داده‌های خام در پایگاه داده‌های بزرگ. هدف اصلی اکتشاف دانش، یافتن دانش نهفته در داده‌ها با کمترین (یا عدم) دخالت انسانی است. مراحل یک پروسه اکتشاف دانش در شکل9-2، نشان داده‌شده است:



شکل9-2: پروسه اکتشاف دانش

گردآوری داده: به طور کلی دو روش برای جمع آوری داده‌ها وجود دارد: در حالت اول خود طراح مدل، تولید داده را نیز کنترل می‌کند. این روش، آزمون طراحی‌شده، نامیده‌می‌شود. دومین وضعیت هنگامی است که شخص خبره نمی‌تواند تأثیری در فرآیند تولید داده داشته باشد. این وضعیت با نام رویکرد مشاهده‌ای شناخته‌می‌شود.

پاک‌سازی داده‌ها (Data Cleaning): در این مرحله داده‌های غیرمعتبر از مجموعه داده‌های آموزشی خارج‌می‌شود. داده‌های دارای خطا یا پرت و اطلاعات ناکامل، نمونه‌هایی از داده‌های زائد هستند که باید پاک‌سازی در مورد آنها انجام شود.

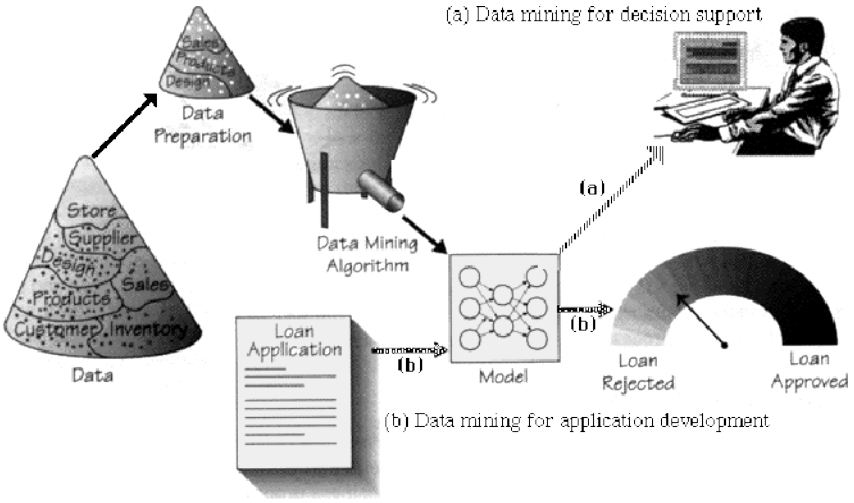
تبدیل داده‌ها (Data Transformation): در این گام داده‌ها به قالبی قابل استفاده برای داده‌کاوی در می‌آیند.

برآورد مدل یا داده کاوی: بخش اصلی فرآیند داده‌کاوی این بخش است که در آن با استفاده از روش‌ها و تکنیک‌های خاص، استخراج الگوهای دانش صورت‌می‌گیرد. به طورکلی روش‌ها و الگوریتم‌های مختلفی جهت یادگیری و تولید یک مدل بر اساس داده‌های ورودی وجوددارد. به نوعی الگوریتم‌های مزبور را می‌توان یک روال جستجو نیز در نظر گرفت. این روال سعی‌می‌کند مدلی پیدا کند که به بهترین نحو داده‌های ورودی را پوشش‌دهد. بایستی توجه نمود که الگوریتم‌های داده‌کاوی که در این مرحله اجرا می‌گردند، با توجه به ماهیت مسأله‌ای که فرآیند داده‌کاوی سعی در تحلیل داده‌های آن را دارد، طراحی می‌گردند. به عبارت دیگر الگوریتم مزبور با توجه به انواع کاربردهای داده‌کاوی، پیاده‌سازی می‌گردد.

ارزیابی الگوها (Pattern Evaluation): تشخیص الگوهای صحیح مورد نظر از سایر الگوها در این مرحله انجام‌می‌شود. صحت الگوها بر اساس یک‌سری از معیارهای سنجیده انجام می‌شود.

نمایش دانش (Knowledge Representation): در این بخش به منظور ارائه‌ی دانش استخراج شده به کاربر، از یک‌سری ابزارهای بصری‌سازی استفاده‌می‌شود. برای استفاده از دانش و مدل استخراج شده، بایستی آن دانش قابل تفسیر باشد. این امر به خاطر آن است که انسان تمایل ندارد که اساس و پایه‌ی تصمیم‌های خود را بر مبنای مدل‌های پیچیده‌ی جعبه سیاه قرار دهد. موضوع مهمی که اینجا وجود دارد آن است که اهداف دقت مدل و قابلیت درک آسان، معمولاً با هم در تضاد هستند. اغلب مدل‌های ساده، قابلیت تفسیر بهتری دارند اما دقت آنها پایین‌تر است. از طرف دیگر مدل‌های دقیق معمولاً ساختار پیچیده‌ای دارند.

پرسش: شکل 9-3 را با توجه به مراحل پروسه اکتشاف دانش تشریح کنید.



شکل9-3: فرایند داده‌کاوی در یک مثال کاربردی

نکته: واژه‌های «داده‌کاوی»، «اکتشاف دانش در پایگاه‌داده» و «استخراج دانش از پایگاه‌داده» اغلب به صورت مترادف با یکدیگر استفاده‌می‌شوند.

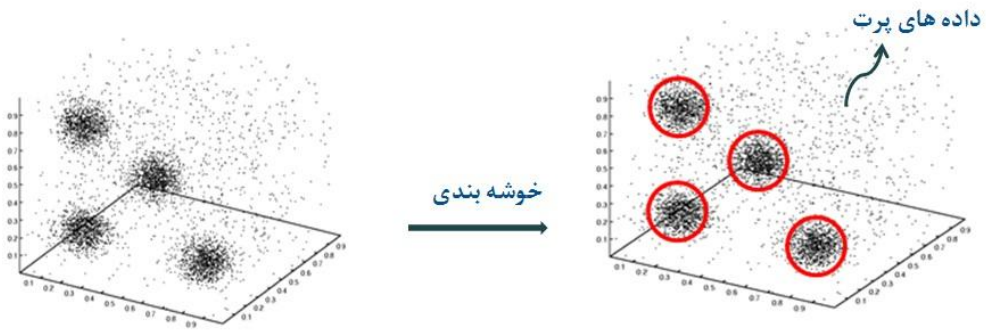
انواع روش‌های اکتشاف دانش(داده‌کاوی) از داده‌ها

- خوشه‌بندی (Clustering): در خوشه‌بندی، هدف یافتن مجموعه متناهی از خوشه‌ها برای توصیف داده‌هاست.
- دسته‌بندی (Classification): هدف در دسته‌بندی داده‌ها این است که یک مدل پیشگویی‌کننده بدست آوریم که این مدل اولا توانایی دسته‌بندی داده‌های ورودی را داشته باشد و ثانیاً بتوان از آن جهت پیشگویی برای تعیین دسته‌ی یک داده که تازه به سیستم اضافه شده، استفاده نمود.
- تخمین (Regression): هدف در رگرسیون ارائه‌ی یک مدل پیشگویی کننده با توانایی نگاشت یک نمونه‌ی داده‌ای به یک متغیر تخمینی است.
- ...

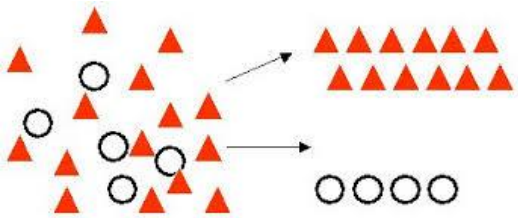
خوشه‌بندی عبارتست از قرار دادن اشیای داده‌ای در چند خوشه به نحوی که:

- داده‌هایی که در یک دسته قرارمی‌گیرند، بیشترین میزان شباهت را به هم داشته باشند.
- داده‌هایی که در دسته‌های مجزا هستند، بیشترین عدم شباهت را به هم داشته باشند.

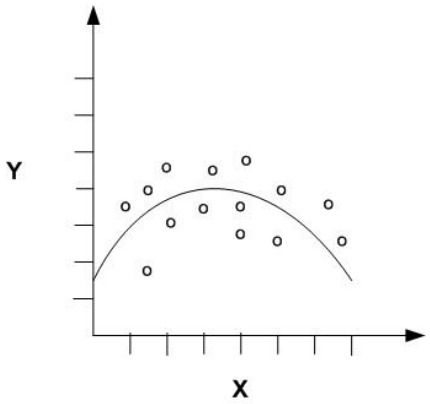
شکل9-4، نحوه عملکرد روش‌های خوشه‌بندی را شرح‌می‌دهد. همانطور که در شکل9-4 مشخص است داده ها به چهار خوشه مجزا تقسیم شده‌اند و داده‌های باقی مانده هم به عنوان داده‌های پرت و نویزی تلقی می‌شوند.



شکل9-4: خوشه‌بندی داده‌ها



شکل9-5: دسته‌بندی داده‌ها



شکل9-6: رگرسیون داده‌ها

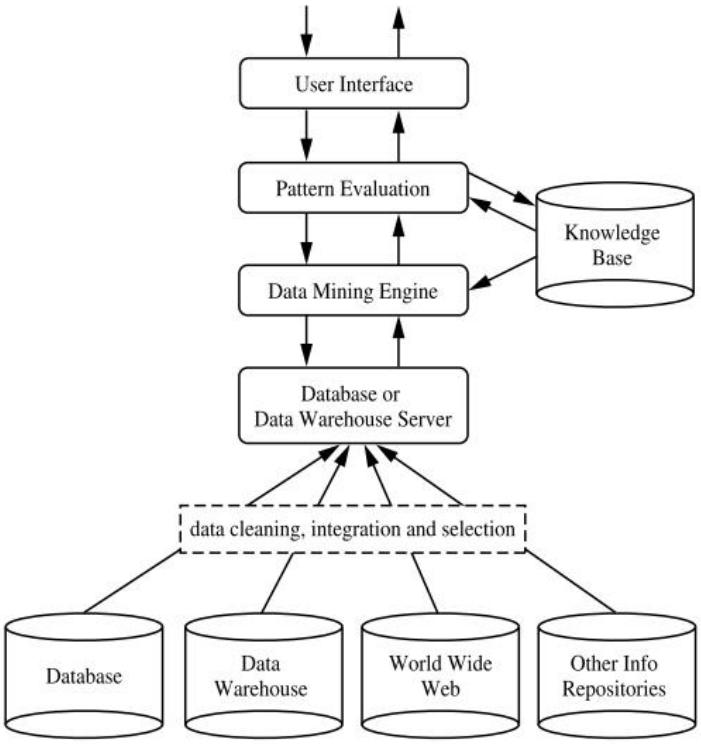
داده‌کاوی چه کارهایی نمی‌تواند انجام دهد؟

داده‌کاوی فقط یک ابزار است و نه یک عصای جادویی. داده‌کاوی به این معنی نیست که شما راحت به کناری بنشینید و ابزارهای داده‌کاوی همه کار را انجام دهد. داده‌کاوی نیاز به شناخت داده‌ها، ابزارهای تحلیل و افراد خبره در این زمینه‌ها را از بین نمی‌برد. داده‌کاوی فقط به تحلیلگران برای پیداکردن الگوها و روابط بین داده‌ها کمک‌می‌کند و در این مورد نیز روابطی که یافته‌می‌شود باید به وسیله داده‌های واقعی، دوباره بررسی و تست گردد.

تمرین‌ها

تمرین9-1*: چه روش‌هایی دیگری علاوه بر روش‌های معرفی‌شده، در داده‌کاوی(استخراج دانش) استفاده‌می‌شوند؟

تمرین9-2: شکل9-7، معماری یک سیستم داده‌کاوی را نشان‌می‌دهد، با مثال کاربردی این معماری را تشریح کنید.



شکل9-7: معماری یک سیستم داده‌کاوی(Han & Kamber, 2006)

تمرین9-3*: روش K نزدیکترین همسایه (KNN = K Nearest Neighbor) چگونه کار می‌کند؟ با مثال ساده، توضیح‌دهید. KNN روش دسته‌بندی محسوب می‌شود یا خوشه‌بندی؟ چرا؟

تمرین9-4*: مراحل الگوریتم K-Means چیست؟ روند کار آن را با مثال توضیح‌دهید. این روش جزء روش دسته‌بندی است یا خوشه‌بندی؟ چرا؟