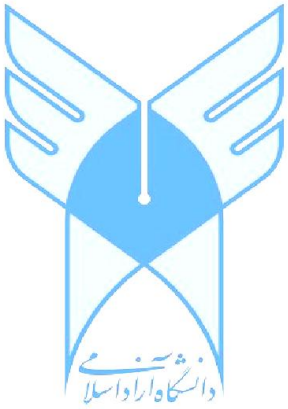


دانشگاه آزاد اسلامی ورزقان

عنوان پروژه

داده کاوی، مفاهیم و کاربرد

پروژه

کارشناسی مهندسی نرم افزار کامپیوتر

استاد راهنما:

جناب آقای مهندس رزمجو

تهیه کننده:

ابراهیم هژبر

۱۳۹۲

فهرست

چکیده

امروزه با گسترش سیستم های پایگاهی و حجم بالای داده ها ی ذخیره شده در این سیستم ها ، نیاز به ابزاری است تا بتوان داده های ذخیره شده را پردازش کردواطلاعات حاصل از این پردازش را در اختیار کاربران قرار داد .

با استفاده از پرسش های ساده در SQL و ابزارهای گوناگون گزارش گیری معمولی ، می توان اطلاعاتی را در اختیار کاربران قرار داد تا بتوانند به نتیجه گیری در مورد داده ها و روابط منطقی میان آنها بپردازند اما وقتی که حجم داده ها بالا باشد ، کاربران هر چند زبر دست و با تجربه باشند نمی توانند الگوهای مفید را در میان حجم انبوه داده ها تشخیص دهند و یا اگر قادر به این کار هم با شند ، هزینه عملیات از نظر نیروی انسانی و مادی بسیار بالا است .

از سوی دیگر کاربران معمولا فرضیه ای را مطرح می کنند و سپس بر اساس گزارشات مشاهده شده به اثبات یا رد فرضیه می پردازند ، در حالی که امروزه نیاز به روشهایی است که اصطلاحا به کشف دانش بپردازند یعنی با کمترین دخالت کاربر و به صورت خودکار الگوها و رابطه های منطقی را بیان نمایند .

داده کاوی یکی از مهمترین این روشها است که به وسیله آن الگوهای مفید در داده ها با حداقل دخالت کاربران شناخته می شوند و اطلاعاتی را در اختیار کاربران و تحلیل گران قرار می دهند تا براساس آنها تصمیمات مهم و حیاتی در سازمانها اتخاذ شوند .

در داده کاوی از بخشی از علم آمار به نام تحلیل اکتشافی داده ها استفاده می شود که در آن بر کشف اطلاعات نهفته و ناشناخته از درون حجم انبوه داده ها تاکید می شود . علاوه بر این داده کاوی با هوش مصنوعی و یادگیری ماشین نیز ارتباط تنگاتنگی دارد ، بنابراین می توان گفت در داده کاوی تئوریهای

پایگاه داده ها ، هوش مصنوعی ، یادگیری ماشین و علم آمار را در هم می آمیزند تا زمینه کاربردی فراهم شود .

باید توجه داشت که اصطلاح داده کاوی زمانی به کار برده می شود که با حجم بزرگی از داده ها ، در حد مگا یا ترابایت ، مواجه باشیم . در تمامی منابع داده کاوی بر این مطلب تاکید شده است .

هر چه حجم داده ها بیشتر و روابط میان آنها پیچیده تر باشد دسترسی به اطلاعات نهفته در میان داده ها مشکلتر می شود و نقش داده کاوی به عنوان یکی از روشهای کشف دانش ، روشن تر می گردد .

مقدمه

با گسترش فناوری اطلاعات و ارتباطات[1] درجهان و ورود سریع آن به زندگی روزمره مردم مسائل و ضرورت‌های تازه‌ای به‌وجودآمده‌است. امروزه انسان توسعه یافته کسی است که به اطلاعات دسترسی داشته‌باشد و دسترسی به اطلاعات نه یک ضرورت،که یک قدرت محسوب‌می‌شود. دراین‌میان شهرها به عنوان مراکز قدرت انسانی و تمدن‌های بشری بیش از پیش اهمیت‌یافته‌اند. به اعتقاد الوین تافلر، مردم کره زمین تا به امروز سه موج اساسی تحول راپشت سرگذاشته اند :

موج اول، موج انقلاب کشاوزی است که زمان آغاز آن برکسی مشخص نیست.

موج دوم، انقلاب صنعتی است که به دنبال اختراع ماشین بخار در سال ۱۷۶۴آغاز شد.

موج سوم یا انقلاب انفورماتیک است که ازسال ۱۹۴۶ که بشر به ساخت کامپیوتر نائل آمده آغاز گشته‌است.

اگر در موج دوم سخت‌افزارها به کمک انسان‌ها می‌آمدند، درموج سوم این نرم‌افزارها هستند که به خدمت بشر می‌شتابند و تفکرات و تصورات آدمی را به شکل کدهای صفر و یک و با کمک امواج ماهواره‌ای مبادله می‌کنند.

در موج سوم، انسان هر روز که بیشتر یاد می‌گیرد، بیشترمی فهمدکه با حقیقت فاصله دارد. موج سوم راموج خردورزی نیز لقب داده اند زیرا در این عرصه‌ها، انسان‌ها دیگر فرصت ندارند زیاد با هم صحبت‌کنند، همه چیز تعریف‌شده و برای هر تعریف، یک کد درنظرگرفته‌شده‌است.

از سوی دیگر در دنیای به شدت رقابتی امروز، اطلاعات بعنوان یکی از فاکتورهای تولیدی مهم پدیدار شده است. در نتیجه تلاش برای استخراج اطلاعات از داده ها توجه بسیاری از افراد دخیل در صنعت

[1] Information and Communication Technology(ICT)

اطلاعات و حوزه های وابسته را به خود جلب نموده است.

حجم بالای داده های دائما در حال رشد در همه حوزه ها و نیز تنوع آنها به شکل داده متنی، اعداد، گرافیکها، نقشه ها، عکسها، تصاویر ماهواره ای و عکسهای گرفته شده با اشعه ایکس نمایانگر پیچیدگی کار تبدیل داده ها به اطلاعات است. علاوه بر این، تفاوت وسیع در فرآیندهای تولید داده مثل روش آنالوگ مبتنی بر کاغذ و روش دیجیتالی مبتنی بر کامپیوتر، مزید بر علت شده است. استراتژیها و فنون متعددی برای گردآوری، ذخیره، سازماندهی و مدیریت کارآمد داده های موجود و رسیدن به نتایج معنی دار بکار گرفته شده اند. بعلاوه، عملکرد مناسب ابرداده که داده ای درباره داده است در عمل عالی بنظر میرسد.

پیشرفتهای حاصله در علم اطلاع رسانی و تکنولوژی اطلاعات، فنون و ابزارهای جدیدی برای غلبه بر رشد مستمر و تنوع بانکهای اطلاعاتی تامین می کنند. این پیشرفتها هم در بعد سخت افزاری و هم نرم افزاری حاصل شده اند.

ریزپردازنده های سریع، ابزارهای ذخیره داده های انبوه پیوسته و غیر پیوسته، اسکنرها، چاپگرها و دیگر ابزارهای جانبی نمایانگر پیشرفتهای حوزه سخت افزار هستند. پیشرفتهای حاصل در نظامهای مدیریت بانک اطلاعات در طی چهار دهه گذشته نمایانگر تلاشهای بخش نرم افزاری است.

این تلاشها در بخش نرم افزار را میتوان بعنوان یک حرکت پیشرونده از ایجاد یک بانک اطلاعات ساده تا شبکه ها و بانکهای اطلاعاتی رابطه ای و سلسله مراتبی برای پاسخگویی به نیاز روزافزون سازماندهی و بازیابی اطلاعات ملاحظه نمود. بدین منظور در هر دوره، نظامهای مدیریت بانک اطلاعاتی مناسب سازگار با نرم افزار سیستم عامل و سخت افزار رایج گسترش یافته اند. در این رابطه میتوان از محصولاتی مانند، Dbase-IV, Unify, Sybase, Oracle و غیره نام برد.

داده کاوی یکی از پیشرفتهای اخیر در راستای فن آوریهای مدیریت داده هاست. داده کاوی مجموعه ای از فنون است که به شخص امکان میدهد تا ورای داده پردازی معمولی حرکت کند و به استخراج اطلاعاتی که در انبوه داده ها مخفی و یا پنهان است کمک می کند. انگیزه برای گسترش داده کاوی بطور عمده از دنیای تجارت در دهه ۱۹۹۰ پدید آمد. مثلا داده کاوی در حوزه بازاریابی، بدلیل پیوستگی غیرقابل انتظاری که بین پروفایل یک مشتری و الگوی خرید او ایجاد میکند اهمیتی خاص دارد.

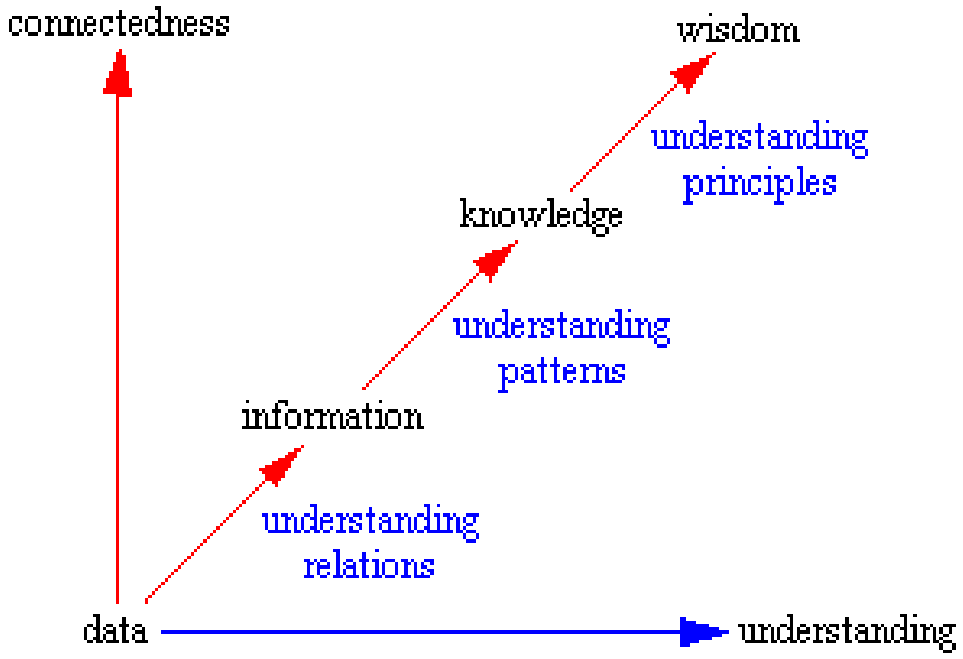
تحلیل رکوردهای حجیم نگهداری سخت افزارهای صنعتی، داده های هواشناسی و دیدن کانال های تلوزیونی از دیگر کاربردهای آن است. در حوزه مدیریت کتابخانه کاربرد داده کاوی بعنوان فرایند ماخذ کاوی نامگذاری شده است. این مقاله به کاربردهای داده کاوی در مدیریت کتابخانه ها و موسسات آموزشی می پردازد. در ابتدا به چند سیستم سازماندهی داده ها که ارتباط نزدیکی به داده کاوی دارند می پردازد؛ سپس عناصر داده ای توصیف میشوند و درپایان چگونگی بکارگیری داده کاوی در کتابخانه ها و موسسات آموزشی مورد بحث قرار گرفته و مسائل عملی مرتبط در نظر گرفته می شوند.

فصل اول – مفاهیم داده کاوی

مدیریت ذخیره سازی و دستیابی اطلاعات

داده های اطلاعاتی به عنوان یکی از منابع حیاتی سازمان شناخته می شود و بسیاری از سازمان ها با اطلاعات و دانش سازمانی خود مانند سایر دارایی های ارزشمندشان برخورد می کنند.

نکته :داده اطلاعاتی (Data) به اطلاعات خام سازمان اتلاق می‌شود و اطلاعات (Information) به داده‌های پردازش شده .همچنین داده های پردازش شده پس از طبقه بندی و آنالیز به دانش سازمان (Knowledge)تبدیل می گردند .

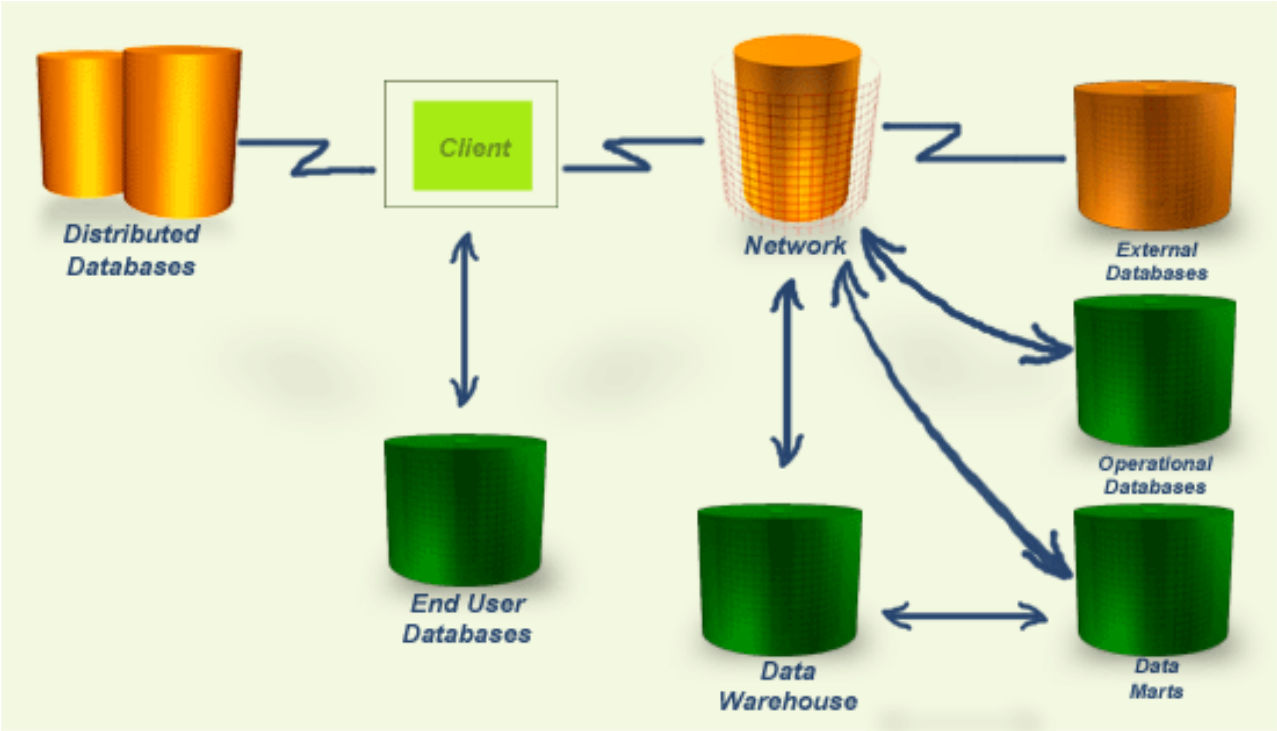


حال تصور نمایید، دسترسی به اطلاعات (Information) در شرایطی که داده‌ها به روش نامناسبی نگهداری شوند و یا روش ضابطه مندی جهت دستیابی به آنها وجود نداشته باشد تا چه حد مشکل

است . برای رسیدن به یک سیستم اطلاعاتی مناسب، داده‌ها می بایست به صورتی منطقی طبقه بندی و ذخیره شوند تا استفاده از آن ها ساده‌تر بوده، با کارایی بیشتری تحلیل شوند و سریعتر مورد استفاده قرار گیرند و در نتیجه مدیریت بهتری بر آن ها اعمال شود.

ساختار بانک اطلاعاتی سازمان :

داده های سازمان ها در انواع بانک های اطلاعاتی و با ساختارهای متنوعی ذخیره می‌گردند . طراحی و سازماندهی این ساختارها، بکارگیری و انتقال به بانک‌های اطلاعاتی پیشرفته و بهینه سازی آن ها یکی خدماتی است که توسط واحدهای فناوری اطلاعات ارایه می شود .

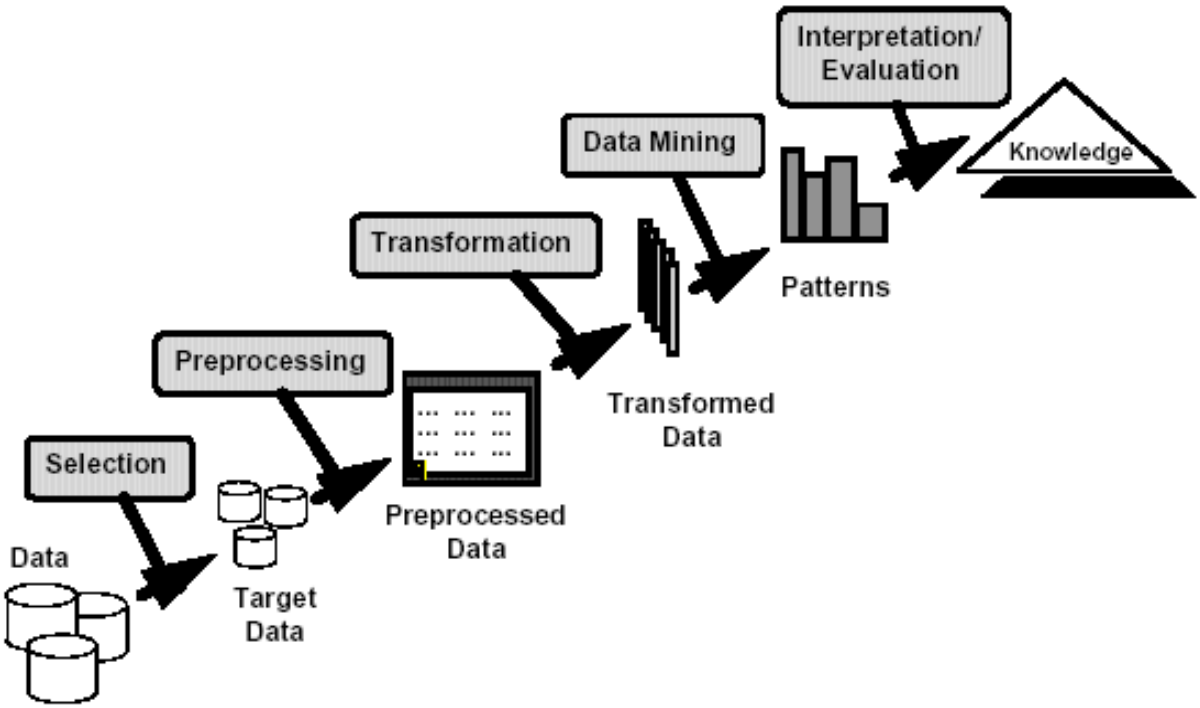


داده کاوی:**(Data Mining)**

داده‌کاوی چیست؟

فناوری‌های نوین اطلاعاتی و ارتباطی، و همچنین تکنولوژی‌های پشتیبان تصمیم، با جمع‌آوری، ذخیره، ارزیابی، تفسیر و تحلیل، بازیابی و اشاعه اطلاعات و دانش به کاربران خاص، می‌توانند در اطلاع‌یابی به‌موقع، صحیح و موردنیاز به افراد تاثیر زیادی داشته‌باشند. یکی از ابزارهای مورد استفاده در این فناوری‌ها، داده کاوی می باشد. داده‌کاوی شامل استفاده از ابزارهای پیشرفته تحلیل داده به منظور کشف الگوهای معتبر، از قبل ناشناخته و روابط در مجموعه داده‌های بزرگ است. این ابزارها، مدلهای

آماری، الگوریتم‌های ریاضی و متدهای یادگیری ماشین[۲] (الگوریتم‌هایی که عملکرد خود را از طریق تجربه به‌صورت اتوماتیک بهبود می‌دهند) می‌باشد. داده‌کاوی فراتر از جمع‌آوری و مدیریت داده است، و شامل تجزیه و تحلیل و پیش‌گویی می‌شود. نام دیگر آن کشف دانش در پایگاه داده یا به اختصار KDD[۳] است .



داده‌کاوی می‌تواند روی داده‌های کمی، متنی، یا چندرسانه‌ای انجام‌گیرد. کاربردهای آن شامل موارد زیر می‌باشد :

- قوانین وابستگی[۴]: الگوهایی که در آن وجود یک آیتم دلالت بر وجود آیتم دیگر دارد،

[۲] Machine Learning

[3] Knowledge Discover in Database

[4] Acossiation Rule

- کلاس‌بندی: انتساب الگوها به یک مجموعه کوچک از کلاس‌های از قبل تعریف شده به وسیله کشف بعضی روابط بین ویژگی‌ها،

- خوشه‌بندی[5]: گروه‌بندی مشتریان یا مجموعه الگوهایی که ویژگی‌های مشابهی دارند،

- پیش‌گویی[6]: کشف الگوها برای پیش‌گویی منطقی درباره آینده،

- تحلیل مسیر[7] یا الگوهای ترتیبی: الگوهایی که در آن یک رخداد منجر به وقوع رخداد دیگر می‌شود.

داده‌کاوی یک تکنولوژی جدید نیست ولی کاربرد آن به‌طور معناداری در بخش‌های مختلف خصوصی و عمومی روبه‌رشد بوده و عموما صنایعی چون بانک، بیمه، پزشکی و خرده‌فروشی از داده-کاوی به هدف کاهش هزینه‌ها، افزایش تحقیقات و افزایش فروش استفاده می‌کنند.

مفاهیم پایه در داده کاوی

در داده کاوی معمولا به کشف الگوهای مفید از میان داده ها اشاره می شود . منظور از الگوی مفید ، مدلی در داده ها است که ارتباط میان یک زیر مجموعه از داده ها را توصیف می کند و معتبر ، ساده ، قابل فهم و جدید است .

[5] Clustering
[6] prediction
[7] Pth Analysis

تعریف داده کاوی

در متون آکادمیک تعاریف گوناگونی برای داده کاوی ارائه شده است . در برخی از این تعاریف داده کاوی در حد ابزاری که کاربران را قادر به ارتباط مستقیم با حجم عظیم داده ها می سازد معرفی گردیده است و در برخی دیگر ، تعاریف دقیقتر که درآنها به کاوش در داده ها توجه می شود موجود است . برخی از این تعاریف عبارتند از :

داده کاوی عبارت است از فرایند استخراج اطلاعات معتبر ، از پیش ناشناخته ، قابل فهم و قابل اعتماد از پایگاه داده های بزرگ و استفاده از آن در تصمیم گیری در فعالیت های تجاری مهم.

اصطلاح داده کاوی به فرایند نیم خودکار تجزیه و تحلیل پایگاه داده های بزرگ به منظور یافتن الگوهای مفید اطلاق می شود

داده کاوی یعنی جستجو در یک پایگاه داده ها برای یافتن الگوهایی میان داده ها .

داده کاوی یعنی استخراج دانش کلان ، قابل استناد و جدید از پایگاه داده ها ی بزرگ .

داده کاوی یعنی تجزیه و تحلیل مجموعه داده های قابل مشاهده برای یافتن روابط مطمئن بین داده ها .

همانگونه که در تعاریف گوناگون داده کاوی مشاهده می شود ، تقریبا در تمامی تعاریف به مفاهیمی چون استخراج دانش ، تحلیل و یافتن الگوی بین داده ها اشاره شده است .

تاریخچه داده کاوی

اخیرا داده کاوی موضوع بسیاری از مقالات ، کنفرانس ها و رساله ها ی عملی شده است ، اما این واژه تا اوایل دهه نود مفهومی نداشت وبه کار برده نمی شد .

در دهه شصت و پیش از آن زمینه هایی برای ایجاد سیستم ها ی جمع آوری و مدیریت داده ها ایجاد شد و تحقیقاتی در این زمینه انجام پذیرفت که منجر به معرفی و ایجاد سیستم های مدیریت پایگاه داده ها گردید .

ایجاد و توسعه مدلهای داده ای برای پایگاه سلسله مراتبی ، شبکه ای و بخصوص رابطه ای در دهه هفتاد ، منجر به معرفی مفاهیمی همچون شاخص گذاری و سازماندهی داده ها و در نهایت ایجاد زبان پرسش SQL در اوایل دهه هشتاد گردید تا کاربران بتوانند گزارشات و فرمهای اطلاعاتی مورد نظر خود را ، از این طریق ایجاد نمایند .

توسعه سیستم های پایگاهی پیشرفته در دهه هشتاد و ایجاد پایگاه های شی گرا ، کاربرد گرا و فعال باعث توسعه همه جانبه و کاربردی شدن این سیستم ها در سراسر جهان گردید . بدین ترتیب DBMS هایی همچون DB2 ، Oracle ، Sybase ، ... ایجاد شدند و حجم زیادی از اطلاعات با استفاده از این سیستم ها مورد پردازش قرار گرفتند . شاید بتوان مهمترین جنبه در معرفی داده کاوی را مبحث کشف دانش از پایگاه داده ها (KDD) دانست بطوری که در بسیاری موارد DM و KDD بصورت مترادف مورد استفاده قرار می گیرند .

برای اولین بار مفهوم داده کاوی در کارگاه IJCAI در زمینه KDD توسط Shapir مطرح گردید . به دنبال آن در سالهای ۱۹۹۱ تا ۱۹۹۴ ، کارگاههای KDD مفاهیم جدیدی را در این شاخه از علم ارائه کردند بطوری که بسیاری از علوم و مفاهیم با آن مرتبط گردیدند.

برخی از کاربردهای داده کاوی در محیطهای واقعی عبارتند از :

خرده فروشی : از کاربردهای کلاسیک داده کاوی است که می توان به موارد زیر اشاره کرد :

تعیین الگوهای خرید مشتریان

تجزیه و تحلیل سبد خرید بازار

پیشگویی میزان خرید مشتریان از طریق پست (فروش الکترونیکی)

بانکداری :

پیش بینی الگوهای کلاهبرداری از طریق کارتهای اعتباری

تشخیص مشتریان ثابت

تعیین میزان استفاده از کارتهای اعتباری بر اساس گروههای اجتماعی

بیمه :

تجزیه و تحلیل دعاوی

پیشگویی میزان خرید بیمه نامه های جدید توسط مشتریان

پزشکی :

تعیین نوع رفتار با بیماران و پیشگویی میزان موفقیت اعمال جراحی

تعیین میزان موفقیت روشهای درمانی در برخورد با بیماریهای سخت

مراحل فرایند کشف دانش از پایگاه داده ها

فرایند کشف دانش از پایگاه داده ها شامل پنج مرحله است که عبارتند از :

انبارش داده ها

انتخاب داده ها

تبدیل داده ها

کاوش در داده ها

تفسیر نتیجه

همانگونه که مشاهده می شود داده کاوی یکی از مراحل این فرایند است که به عنوان بخش چهارم آن نقش مهمی در کشف دانش از داده ها ایفا می کند .

انبارش داده ها

وجود اطلاعات صحیح و منسجم یکی از ملزوماتی است که در داده کاوی به آن نیازمندیم . اشتباه و عدم وجود اطلاعات صحیح باعث نتیجه گیری غلط و در نتیجه اخذ تصمیمات ناصحیح در سازمانها می گردد و منتج به نتایج خطرناکی خواهد گردید که نمونه های آن کم نیستند .

اکثر سازمانها دچار یک خلا اطلاعاتی هستند . در اینگونه سازمانها معمولا سیستم های اطلاعاتی در طول زمان و با معماری و مدیریت های گوناگون ساخته شده اند ، به طوری که سازمان اطلاعاتی یکپارچه و مشخصی مشاهده نمی گردد . علاوه بر این برای فرایند داده کاوی به اطلاعات خلاصه و مهم در زمینه تصمیم گیریهای حیاتی نیازمندیم .

هدف از فرایند انبارش داده ها فراهم کردن یک محیط یکپارچه جهت پردازش اطلاعات است . در این فرایند ، اطلاعات تحلیلی و موجز در دوره های مناسب زمانی سازماندهی و ذخیره می شود تا بتوان از آنها در فرایند های تصمیم گیری که از ملزومات آن داده کاوی است ، استفاده شود . به طور کلی تعریف زیر برای انبار داده ها ارائه می گردد :

انبار داده ها ، مجموعه ای است موضوعی، مجتمع ، متغیر در زمان و پایدار از داده ها که به منظور پشتیبانی از فرایند مدیریت تصمیم گیری مورد استفاده قرار می گیرد .

انبارش داده ها خود موضوع مفصلی است که مقاله ها و رساله ها ی گوناگونی در مورد آن نگاشته شده اند . در این فصل به منظور آشنایی با این فرایند به آن اشاره ای شد .

انتخاب داده ها

انبار داده ها شامل انواع مختلف و گوناگونی از داده ها است که همه آنها در داده کاوی مورد نیاز نیستند . برای فرایند داده کاوی باید داده ها ی مورد نیاز انتخاب شوند . به عنوان مثال در یک پایگاه داده های مربوط به سیستم فروشگاهی ، اطلاعاتی در مورد خرید مشتریان ، خصوصیات آماری آنها ، تامین کنندگان ، خرید ، حسابداری و ... وجود دارند . برای تعیین نحوه چیدن قفسه ها تنها به داده ها یی در مورد خرید مشتریان و خصوصیات آماری آنها نیاز است . حتی در مواردی نیاز به کاوش در تمام محتویات پایگاه نیست بلکه ممکن است به منظور کاهش هزینه عملیات ، نمونه هایی از عناصر انتخاب و کاوش شوند .

تبدیل داده ها

هنگامی که داده های مورد نیاز انتخاب شدند و داده ها ی مورد کاوش مشخص گردیدند ، معمولا به تبدیلات خاصی روی داده ها نیاز است . نوع تبدیل به عملیات و تکنیک داده کاوی مورد استفاده بستگی دارد : تبدیلاتی ساده همچون تبدیل نوع داده ای به نوع دیگر تا تبدیلات پیچیده تر همچون تعریف صفات جدید با انجام عملیاتهای ریاضی و منطقی روی صفات موجود .

کاوش در داده ها

داده های تبدیل شده با استفاده از تکنیکها و عملیاتهای داده کاوی مورد کاوش قرار می گیرند تا الگوهای مورد نظر کشف شوند .

تفسیر نتیجه

اطلاعات استخراج شده با توجه به هدف کاربر تجزیه و تحلیل و بهترین نتایج معین می گردند . هدف از این مرحله تنها ارائه نتیجه (بصورت منطقی و یا نموداری) نیست ، بلکه پالایش اطلاعات ارایه شده به کاربر نیز از اهداف مهم این مرحله است .

عملیاتهای داده کاوی

در داده کاوی ، چهار عمل اصلی انجام می شود که عبارتند از

مدلسازی پیشگویی کننده

تقطیع پایگاه داده ها

تحلیل پیوند

تشخیص انحراف

از عملیاتهای اصلی مذکور ، یک یا بیش از یکی از آنها در پیاده سازی کاربرد های گوناگون داده کاوی استفاده می شوند . به عنوان مثال برای کاربرد های خرده فروشی معمولا از عملیات تقطیع و تحلیل پیوند استفاده می شود در حالی که برای تشخیص کلاهبرداری ، می توان از هر یک از چهار عملیات مذکور استفاده نمود . علاوه برا ین می توان از دنباله ای از عملیاتها برای یک منظور خاص استفاده کرد . مثلا برای شناسایی مشتریان ، ابتدا پایگاه تقطیع می شود و سپس مدلسازی پیشگویی کننده در قطعات ایجاد شده اعمال می گردد .

تکنیکها ، روشها و الگوریتمهای داده کاوی ، راههای پیاده سازی عملیاتهای داده کاوی هستند . اگر چه هر عملیات نقاط ضعف و قوت خود را دارد ، ابزارهای گوناگون داده کاوی عملیاتها را بر اساس معیارهای خاصی ، انتخاب می کنند . این معیارها عبارتند از :

تناسب با نوع داده های ورودی

شفافیت خروجی داده کاوی

مقاومت در مقابل اشتباه در مقادیر داده ها

میزان صحت خروجی

توانایی کار کردن با حجم بالای داده ها

مدلسازی پیشگویی کننده

مدلسازی پیشگویی کننده ، شبیه تجربه یادگیری انسان در به کار بردن مشاهدات برای ایجاد یک مدل از خصوصیات مهم پدیده ها است . در این روش از تعمیم دنیای واقعی و قابلیت تطبیق داده های جدید با یک قالب کلی ، استفاده می شود .

در این مدل ، می توان با تحلیل یک پایگاه داده های موجود ، خصوصیات مجموعه های داده را تعیین کرد . این مدل با استفاده از روش یادگیری نظارت شده، شامل دو فاز آموزش و آزمایش ایجاد شده است . در فاز آموزش با استفاده از نمونه های عظیمی از داده های سابقه ای ، مدلی ساخته می شود که به آن مجموعه آموزشی می گویند . در فاز آزمایش این مدل روی داده هایی که در مجموعه آموزشی قرار ندارند ، اعمال می شود تا صحت و خصو صیات آن تایید گردد .

از کاربردهای عمده این مدل می توان به مدیریت مشتریان ، تصویب اعتبار ، بازاریابی مستقیم در خرده فروشی و ... اشاره کرد .

تقطیع پایگاه داده ها

هدف از تقطیع پایگاه داده ها ، تقسیم آن به تعداد نامعینی از قطعات یا خوشه هایی از رکوردهای مشابه است ، یعنی رکوردهایی که خصوصیاتی مشابه دارند و می توان آنها را همگن فرض کرد . پیوستگی داخلی این قطعات بسیار زیاد است در حالی که همبستگی خارجی میان آنها کم می باشد .

در این مدل بر خلاف مدل قبل ، از یادگیری نظارت نشده برای تعیین زیرشاخه های ممکن از جمعیت داده ای استفاده می شود . دقت تقطیع پایگاه داده ها از روشهای دیگر کمتر است ، بنابراین در مقابل خصوصیات نامربوط و افزونگی ، حساسیت کمتری از خود نشان می دهد .

از کاربردهای این روش می توان به شناسایی مشتریان ، بازاریابی مستقیم و ... اشاره کرد .

تحلیل پیوند

در این روش پیوند هایی مرسوم به بستگی میان رکوردها و یا مجموعه ای از رکوردها بازشناسی می شوند . سه رده ویژه از تحلیل پیوند وجود دارند که عبارتند از :

کشف بستگی

کشف الگوهای متوالی

کشف دنباله های زمانی مشابه

تشخیص انحراف

داده کاوی فرآیندی است که طی آن با استفاده از انواع مختلف ابزار تحلیل داده به دنبال کشف الگوها و ارتباطات میان داده های موجود که ممکن است منجر به استخراج اطلاعات جدیدی از پایگاه داده گردند می باشد.

اولین وساده ترین گام تحلیل داده در داده کاوی توضیح و شرح مشخص داده (از جمله معنی داده وانحراف استاندارد کلمه)می باشد که این کار می تواند به وسیله نمودارها و گراف ها وهمچنین کلماتی که با این کلمه ارتباط معنایی نزدیکی دارند انجام گردد در نتیجه جمع آوری، جستجو و انتخاب داده درست در این بخش بسیار مهم و حیاتی می باشد.

اما این کار به تنهایی کار خاصی انجام نمی دهد شما باید یک مدل پیش بینی کننده بر اساس الگوهایی که از نتایج دانش به دست آورده شده بسازید سپس آزمایش کنید که آیا آن مدل با نمونه اصلی سازگار است. یک مدل خوب نباید با جهان واقع تفاوت چندانی داشته باشد

آخرین گام نیز تشخیص صحت وسقم عملکرد مدل بصورت تجربی می باشد. برای مثال از یک بانک مربوط به مشتریان وپاسخ هایی که به یک پیشنهاد خاص داده اند یک مدل می سازید که بر اساس آن مشخص می شود که کدام حدس وانتظار بیشترین نزدیکی را با یک پیشنهاد مانند پیشنهاد قبلی دارد و اینکه آیا شما می توانید بر این حدس اعتماد کنید یا نه؟

الگوریتم های داده کاوی

حال بیایید برخی از الگوریتمها و مدلهایی را که برای کاوش داده استفاده می شود را بررسی کنیم.

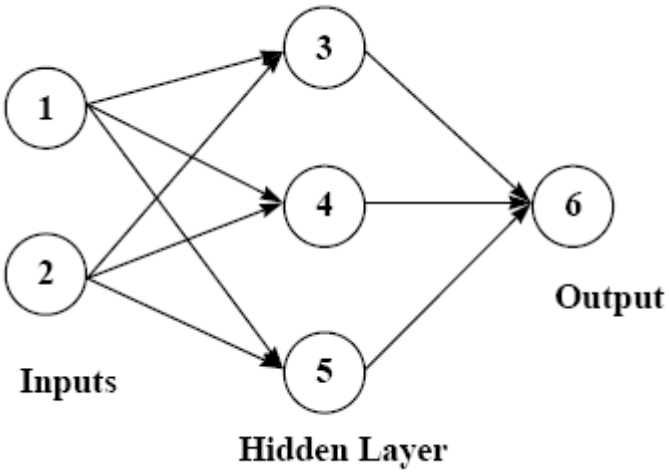
اغلب محصولات از انواع گوناگونی از الگوریتمها که در علم کامپیوتر یا مقالات آماری ارائه شده به همراه پیاده سازی خاص آنها که جهت رسیدن به هدف فروشنده می باشد استفاده می نمایند. برای مثال بسیاری از فروشندگان نسخه هایی از درختهای تصمیم CART یا CHAID را به همراه امکاناتی برای کار بر روی کامپیوترهای موازی می فروشند. برخی از فروشندگان الگوریتمهای مختص خود دارند که گرچه ممکن است وابستگی ها یا امکانات اضافی نداشته باشد اما می تواند خوب کار کند.

شاید مهمترین نکته ای باشد که هیچ مدل یا الگوریتمی نمی تواند و نباید به تنهایی استفاده شود. برای هر مساله داده شده طبیعت داده استفاده شده بر روی انتخاب مدلها و الگوریتمهایی که شما بر می گزینید تاثیر خواهد گذاشت. نمی توان هیچ مدل یا الگوریتمی را در این زمینه بهترین نامید. نتیجتا شما به یک سری ابزار و تکنولوژی جهت یافتن بهترین مدل ممکنه نیاز خواهید داشت.

شبکه های عصبی

شبکه های عصبی به طور خاصی مورد استفاده اند چرا که آنها ابزاری موثر برای مدلسازی مسائل بزرگ و پیچیده که ممکن است در آنها صدها متغیر پیش بینی کننده که فعل و انفعالات زیادی دارند وجود داشته باشد.(شبکه های عصبی زیستی بطور غیر قابل مقایسه ای پیچیده تر هستند.)شبکه های عصبی می توانند در مسائل طبقه بندی یا حدسهای بازگشتی(که در آنها متغیر خروجی پیوسته است) استفاده شوند.

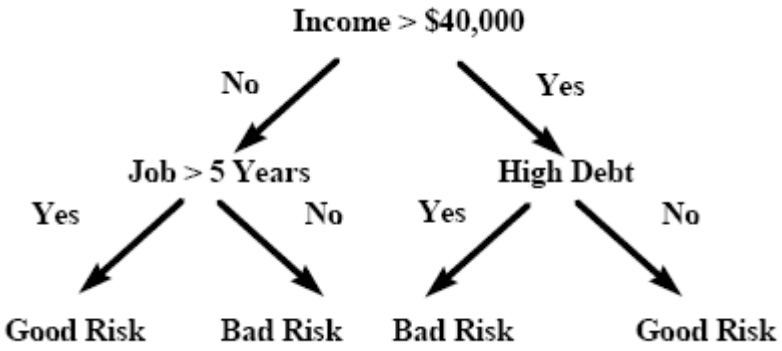
یک شبکه عصبی با یک لایه داخلی شروع می شود که در آن هر گره به یک متغیر پیشگو منسوب می گردد. این گره های ورودی به یک تعداد از گره ها در لایه پنهان متصل می شوند.گره ها در لایه پنهان می توانند به گره هایی در یک لایه پنهان دیگر یا به یک لایه خروجی متصل شود. لایه خروجی خود شامل یک یا بیشتر متغیرهای جواب می باشد.



یک شبکه عصبی با یک لایه پنهان

درخت های انتخاب

درخت های انتخاب راهی برای نمایش یک سری از قوانین که به یک کلاس یا مقدار منجر می شود می باشند. برای مثال شما ممکن است بخواهید درخواستهای وام را برحسب ریسک اعتبار خوب یا بد طبقه بندی کنید. شکل بعد یک مدل ساده از یک درخت انتخاب به همراه توضیح در مورد تمام بسته های پایه آن یعنی گره انتخاب, شاخه ها و برگهای آن که این مساله را حل می کند نشان می دهد.



اولین بسته گره بالایی تصمیم یا ریشه می باشد که یک بررسی جهت برقراری شرط خاصی می نماید. گره ریشه در این مثال "Income>$40,000" می باشد. نتایج این بررسی منجر می شود که درخت به دوشاخه تقسیم گرددکه هر یک نشان دهنده جوابهای ممکن است.در این مورد بررسی شرط مذکور می تواند دارای جواب خیر یا بله باشد در نتیجه دو شاخه داریم.

براساس نوع الگوریتم هر گره می تواند دو یا تعداد بیشتری شاخه داشته باشد. برای مثال CART درختهایی با تنها دوشاخه در هر گره تولید می کند.چنین درختی یک درخت دودویی می باشد.

مدلهای مختلف درخت تصمیم بطور عمومی در داده کاوی برای کاوش داده و برای استنتاج درخت و قوانین آن که برای پیش بینی مورد استفاده قرار می گیرد استفاده می شوند. یک تعداد از الگوریتمهای مختلف می توانند برای ساخت درختهای تصمیم شامل Quest, CART,CHAID و C5.0 بکار روند. اندازه درخت می تواند از طریق قوانین متوقف شونده که رشد درخت را محدود می کنند کنترل شود.

استنتاج قانون

استنتاج قانون روشی برای بدست آوردن یک سری از قوانین برای طبقه بندی موارد می باشد. اگرچه درختهای تصمیم می توانند یک سری قوانین تولید کنند روشهای استنتاج قانون یک مجموعه از قوانین وابسته که ضرورتا درختی تشکیل نمی دهند را تولید می نماید. چون استنتاج کننده قوانین لزوما انشعابی در هر سطح قرار نمی دهد و می تواند گام بعدی را تشخیص دهد گاهی اوقات می تواند الگوهای مختلف و بهتری را برای طبقه بندی بیابد. برخلاف درختان قوانین تولیدی ممکن است تمام حالتهای ممکن را پوشش ندهند.

الگوریتمهای ژنتیک

الگوریتمهای ژنتیک برای یافت الگوها استفاده نمی شود بلکه بیشتر به منظور راهنمایی در مورد فرآیند یادگیری الگوریتمهای داده کاوی مانند شبکه های عصبی مورد استفاده قرار می گیرد. الگوریتمهای ژنتیک به عنوان یک متد جهت انجام یک جستجوی هدایت شده برای مدلهای خوب در فضای حل مساله عمل می کند.

این الگوریتمها, الگوریتمهای ژنتیک نامیده می شوند چون بطور بی قاعده ای الگوی تکامل زیستی که در آن اعضای یک نسل بر سر انتقال خصوصیات خود به نسل بعد رقابت می کنند تا نهایتا بهترین مدل یافت شود را دنبال می کنند. اطلاعاتی که باید انتقال داده شود در قالب کروموزمها که شامل پارامترهایی برای ساختن مدل می باشد قرار می گیرد.

مدل فرآیند دو سویه

مدل فرآیند دو سویه که در زیر توضیح داده شده است برخی از موارد پیش بینی را از مدل CRISP-DM به ارث می برد.

گامهای اصلی داده کاوی جهت کشف دانش عبارتند از:

- تعریف مساله
- ساختن پایگاه داده مربوط به داده کاوی
- جستجوی داده
- آماده ساختن داده برای مدل سازی
- ساختن مدل
- ارزیابی مدل

- ساخت مدل ونتایج

به سراغ این گامها می رویم تا فرآیند کشف دانش را بهتر متوجه شویم.

تعریف مساله

در ابتدای امر پیش زمینه کشف دانش فهم درست داده و مساله می باشد. بدون این فهم درست هیچ الگوریتمی صرف نظر از خبره بودن آن نمی تواند نتیجه مطمئنی برای شما حاصل نماید و همچنین شما قادر نخواهید بود که مسائلی را که سعی در حل آن دارید تعریف کرده و همچنین داده را جهت کاوش آماده نموده و یا نتایج را به طور صحیح تفسیر نمائید. برای استفاده بهتر از داده کاوی شما باید یک بیان واضح از هدف خود داشته باشید.

ساختن یک پایگاه داده داده کاوی

این گام به همراه دو گام بعدی هسته آماده سازی داده را تشکیل می دهند. در مجموع گامهای گفته شده وقت و کار بیشتری از سایر گامها می برند. ممکن است شما گامهای تکراری در آماده سازی داده و ساختن مدل داشته باشید چرا که در هر مرحله ممکن است به نکته ای برسید که شما را بر آن دارد داده خود را بهبود بخشید. این گامهای آماده سازی داده می تواند ۵۰٪ تا ۹۰٪ وقت و کار از تمام فرآیند کشف دانش را به خود اختصاص دهد.

داده ای که می خواهد کاوش شود باید در یک پایگاه داده ذخیره شود. بر اساس مقدار داده, پیچیدگی داده و استفاده هایی که قرار است از آن شود یک فایل معمولی و یا یک SpreadSheet برای این کار کافی است.

به احتمال زیاد شما می خواهید داده موجود در انباره داده را تغییر دهید. به علاوه شما ممکن است بخواهید فیلدهای جدیدی که از فیلدهای موجود محاسبه شده است را به انبار داده خود بیافزایید.این یکی از دلایل استفاده از یک پایگاه داده جداگانه است.

دلیل دیگر برای این کار آن است که انبار داده های یکی شده ممکن است به آسانی انواع جستجوهایی را که شما برای فهم داده به آنها نیاز دارید انجام ندهد. مانند پرس و جوهایی که داده را خلاصه می کند, گزارشات چند بعدی و بسیاری از انواع دیگر از گرافها یا مصورات.

و دلیل آخر اینکه شما ممکن است بخواهید این داده را در یک سیستم مدیریت پایگاه داده به همراه یک طراحی فیزیکی متفاوت از انبار داده خود ذخیره کنید. مردم به طور روز افزونی در حال انتخاب پایگاه داده های خاص منظوره ای هستند که این نیازهای داده کاوی را به نحو مناسبی حمایت کند. به هرحال اگر داده موجود در انبار داده شما اجازه می دهد که مراکز منطقی داده ای ایجادکنید و اگر شما می توانید تقاضای داده کاوی را ارضا نمایید پایگاه داده شما به خوبی وظیفه خود را انجام می دهد.

مراحل لازم برای ساخت یک پایگاه داده داده کاوی به شکل زیر می باشد:

- جمع آوری داده ها
- توضیح داده ها
- انتخاب داده ها
- تعیین کیفیت داده ها و پاک کردن آن
- تثبیت و یکپارچگی
- ساختن فوق داده (داده هایی که خود بیانگر توضیحی در مورد داده های موجود می باشند.)

- بارکردن پایگاه داده مربوط به داده کاوی
- نگهداری پایگاه داده مربوط به داده کاوی

این کارها ممکن است لزوما به همین ترتیب گفته شده انجام نگردند.

جستجوی داده

به بخش توضیح داده برای داده کاوی که توضیح مختصری راجع به اشکال, تجزیه و تحلیل ارتباط و دیگر وسایل جستجوی داده می باشد نگاهی بیاندازید.

هدف شناسایی مهمترین فیلدها در پیش بینی نتیجه و تعیین اینکه کدام یک از داده های بدست آمده مفید می باشد است.

در یک مجموعه داده ای با صدها یا حتی هزاران ستون جستجوی داده می تواند کار و زمان بر باشد. یک واسط مناسب و جواب کامپیوتر سریع در این فاز مهم و حیاتی می باشند زیرا هنگامی که شما برای دریافت پاسخ برخی گراف ها مجبور باشید ۲۰ دقیقه صبر کنید ماهیت جستجوی شما به کلی تغییر خواهد کرد.

آماده سازی داده برای مدل سازی

این آخرین گام آماده سازی داده قبل از ساخت مدلهاست. چهار قسمت مهم در این مرحله وجود دارد:

انتخاب متغیرها

انتخاب سطرها

ساختن متغیرهای جدید

تغییر شکل متغیرها

ساختن مدل داده کاوی

مهمترین مساله برای یادآوری در مورد ساخت مدل آن است که این کار یک فرآیند تکراری است. شما برای جستجو به مدلهای جایگزین جهت یافتن سودمندترین آنها جهت حل مسائلتان نیاز دارید. آنچه که شما در جستجوی یک مدل مناسب یاد می گیرید می تواند شما را به بازگشتن به عقب و انجام برخی تغییرات در داده مورد استفاده خود و حتی بهبود بیان ساله راهنمایی کند.

هنگامی که شما در مورد نوع پیش بینی که می خواهید انجام دهید تصمیم گرفتید باید یک نوع مدل برای ساخت تصمیم خود انتخاب کنید.

آماده سازی و آزمایش مدل داده کاوی احتیاج به این دارد که داده به حداقل دو گروه شکسته شود: یکی برای آماده کردن مدل و دیگری جهت تست مدل مربوطه. اگر شما از آماده سازی و تست متفاوتی استفاده ننمائید دقت مدل خواهد بود.

تائید اعتبارساده

پایه ای ترین روش تست داده تایید اعتبار ساده می باشد. برای انجام این کار چون درصدی از پایگاه داده را به عنوان یک تست پایگاه داده کنار بگذارید و به هر صورت از آن در برآورد و ساخت مدل استفاده ننمائید. این درصد معمولا بین ۵ تا ۳۳ می باشد.

ارزیابی و تفسیر؛ تایید اعتبار مدل

بعد از ساخت یک مدل شما باید نتایج آن را ارزیابی نموده و همچنین اهمیت آن را نیز توضیح دهید.

ایجاد معماری مدل و نتایج

هنگامی که یک مدل ساخته و تایید اعتبار می شود می تواند در دو راه اصلی مورد استفاده قرار گیرد. راه اول برای تحلیل گر است که اعمالی را بر اساس دید ساده از مدل و نتایج آن معرفی می کند. راه دوم بکاربردن مدلها در مجموعه داده ای مختلف است. این مدل می تواند برای مشخص نمودن رکوردها بر اساس گروه بندیشان و یا مقدار دهی یک امتیاز مثلا احتمال انجام یک عمل استفاده گردد. هنگام به دست آوردن یک کاربرد پیچیده داده کاوی اغلب اگر چه بخش بحرانی اما کوچک پروژه نهایی به حساب می آید. برای مثال دانشی که از داده کاوی کشف می شود می تواند با دانش متخصصان داده و تراکنشهای ورودی ترکیب شود. در یک سیستم تشخیص فرآیند الگوهای موجود فرآیند می توانند با الگوهای کشف شده تلفیق شوند. هنگامی که موارد مفروض این فرآیند برای ارزیابی به بررسی کنندگان فرستاده می شوند بررسی کنندگان ممکن است نیاز داشته باشند که به رکوردهایی در پایگاه داده که مربوط به قسمتهای ادعا شده توسط یک سازنده است دسترسی پیدا کنند.

به طور کلی مراحلی که توضیح داده شد برای انجام هر فرآیند داده کاوی لازم به نظر می رسد.

امروزه با گسترش بانکهای اطلاعاتی و حجم عظیم داده های ذخیره شده در این سیستمها، نیاز به ابزاری است که این داده های ذخیره شده را پردازش کند و تبدیل به یک سری اطلاعات مفید و سودمند کند که بتوان با توجه به این اطلاعات، تصمیمات مهم و حیاتی در سازمانها اتخاذ کرد تا به سود بیشتری دست یابند. بنابراین داده کاوی یک سری ابزار در اختیار دارد که به صورت نیمه خودکار و با حداقل دخالت کاربران اطلاعات سودمند و در اصطلاح الگوهای مفید (روابط منطقی بین داده ها) را از میان حجم انبوه داده ها کشف میکند.

از کاربردهای مهم داده کاوی می توان به خرده فروشی، بیمه، بانکها، و ... اشاره کرد.

در فرایند داده کاوی از مدلها و الگوریتم هایی همانند: شبکه های عصبی، درختهای انتخاب، استنتاج قانون و الگوریتمهای ژنتیک استفاده می شود که با استفاده از تکنیکهایی همچون مدلسازی پیشگویی کننده، تقطیع پایگاه داده ها، تحلیل پیوند و تشخیص انحراف می توانیم الگوهای مفید در داده ها را با حداقل دخالت کاربر کشف کنیم.

در نتیجه هدف اصلی در داده کاوی کشف دانش نهفته در داده هاست که در بانکهای عظیم اطلاعاتی وجود دارند که برای دست یافتن به این دانش عظیم بایستی در ابتدا یک محیط یکپارچه از داده ها که پایگاه داده کاوی نامیده می شود فراهم شود سپس داده های مورد نظر جستجو شود آنگاه تبدیلاتی روی آنها صورت گیرد و در مرحله چهارم اکتشاف دانش که داده کاوی نامیده می شود با ابزارهای مورد استفاده در داده کاوی الگوهای موردنظر کشف گردد و در نهایت در مرحله آخر کشف دانش نتیجه به صورت کاملا قابل فهم به کاربر ارائه گردد.

سابقه داده کاوی

داده کاوی و کشف دانش در پایگاه داده ها از جمله موضوع هایی هستند که همزمان با ایجاد و استفاده از پایگاه داده ها در اوایل دهه ۸۰ برای جستجوی دانش در داده ها شکل گرفت.

شاید بتوان لوول (۱۹۸۳) را اولین شخصی دانست که گزارشی در مورد داده کاوی تحت عنوان « شبیه سازی فعالیت داده کاوی » ارائه نمود. همزمان با او پژوهشگران و متخصصان علوم رایانه، آمار، هوش مصنوعی، یادگیری ماشین و . . . نیز به پژوهش در این زمینه و زمینه های مرتبط با آن پرداخته اند.

پژوهش جدی روی موضوع داده کاوی از اوایل دهه ۹۰ شروع شد. پژوهش ها و مطالعه های زیادی در این زمینه صورت گرفته، همچنین سمینارها، دوره های آموزشی و کنفرانس هایی نیز برگزار شده است. نتایج پایه های نظری داده کاوی در تعدادی از مقاله های پژوهشی آورده شده است. مثلاً سال ۱۹۹۱ پیاتتسکی و شاپیرو [۸] « استقلال آماری قاعده ها در داده کاوی» را بررسی نموده اند. سال ۱۹۹۵ هافمن و نش استفاده از داده کاوی و داده انبار[۹] توسط بانک های آمریکا را بررسی نموده و بیان کردند که چگونه این سیستم ها برای بانک های آمریکا قدرت رقابت بیشتری ایجاد می کنند. چت فیلد مشکلات ایجاد شده توسط داده کاوی را بررسی نمود و همچنین مقاله ای تحت عنوان « مدل های خطی غیر دقیق داده کاوی و استنباط آماری » ارایه نمود. هندری نیز دیدگاه اقتصاد سنجی روی داده کاوی را تهیه کرد. در این سال انجمن داده کاوی همزمان با اولین کنفرانس بین المللی «کشف دانش و داده کاوی» شروع به کار کرد. این کنفرانس توسعه یافته چهار دوره آموزشی بین المللی در پایگاه های داده درسال ۱۹۸۹ تا ۱۹۹۴ بود. انجمن مذکور، یک سازمان علمی به نام ACM- SIGKDD را ایجاد نمود. سال ۱۹۹۶ ایمیلنسکی[۱۰] و منیلا[۱۱] دیدگاهی از داده کاوی به عنوان «پرس و جو کننده از پایگاه های استنتاجی[۱۲]» را پیشنهاد کردند. فایاد، پیاتتسکی

---

[8] - piatetsky-shapiro

[۹]- Data warehouse

[۱۰]- Imielnski

[11] - Mannila

[12] - Inductive databases

– شاپیرو، اودورأسامی پیشرفت های کشف دانش و داده کاوی را عنوان کردند. در سال ۱۹۹۷ منیلا خلاصه ای از مطالعه روی اساس داده کاوی ارایه نمود. باربارا و همکاران نیز دیدگاه کاهش داده ها روی داده کاوی را در گزارش کاهش داده های نیوجرسی ارایه نمودند. همچنین می توان برای کاربرد داده کاوی

در مدیریت مالی می توان، تحلیل داده های مالی و مدل سازی مالی بنینگاه و چاچ کز و هیگینز [۱۳] را ملاحظه کرد فریدمن نیز مقاله ای در ارتباط با مفهوم آمار و داده کاوی ارایه نمود. سال ۱۹۹۸ هند [۱۴] مقاله ای تحت عنوان « داده کاوی : آمار یا بیشتر؟ » ارائه نمود. کلینبرگ[۱۵] پائودیمیتریو و راغان [۱۶] دیدگاه اقتصاد سنجی روی داده کاوی و عملکرد داده کاوی به عنوان یک مسئله بهینه را ارایه نمودند. در این سال نیز کنفرانس های ناحیه ای و بین المللی در مورد داده کاوی برگزار شد که از جمله می توان به کنفرانس آسیا و اقیانوسیه درباره کشف دانش و داده کاوی اشاره کرد. سال ۲۰۰۰ هند و همکاران و اسمیت بحث های مقایسه ای بین آمار و داده کاوی را ارایه کردند. سری و استاوا، کولی، رش پاند و تن استفاده از وب در کاوش داده ها و کاربردهای آن را ارایه کردند. سال ۲۰۰۲ کلادیو کانورسانو و همکاران « مدل آمیخته چندگانه جمع پذیر تعمیم یافته » برای داده کاوی را بررسی نمودند. پائلو و گیانلوکاپاسرون، « داده کاوی ساختارهای پیوند برای مدل رفتار مصرف کننده » را ارایه نمودند.

مفهوم داده کاوی

عبارت داده کاوی مترادف با یکی از عبارت های استخراج دانش، برداشت اطلاعات، وارسی داده ها و حتی لایروبی کردن داده هاست که در حقیقت کشف دانش در پایگاه داده ها[۱۷] (**KDD**) را توصیف می کند. بنابراین ایده ای که مبنای داده کاوی است یک فرآیند با اهمیت از شناخت الگوهای بالقوه مفید، تازه و درنهایت قابل درک در داده هاست. واژه کشف دانش در پایگاه داده ها در اوایل دهه ۸۰ در مراجعه به مفهوم کلی، گسترده، سطح بالا و به دنبال جستجوی دانش در اطلاعات

---

13 - Benninga, Czaczkes, Higgins
14 - Hand
15 - Kleinberg
16 - Paodimitriou , Raghavan
17 - Knowledge Discovery of Database

شکل گرفته است. داده کاوی کاربرد سطح بالای فنون و ابزار بکار برده شده برای معرفی و تحلیل داده ها ی تصمیم گیرندگان است. اصطلاح داده کاوی را آمار شناسان، تحلیل گران داده ها و انجمن سیستم های اطلاعات مدیریت به کار برده اند در حالی که پژوهشگران یادگیری ماشین و هوش مصنوعی از **KDD** بیشتر استفاده می کنند. در ادامه چند تعریف از داده کاوی ارائه می شود.

۱- «داده کاوی یا به تعبیر دیگر کشف دانش در پایگاه داده ها، استخراج غیر بدیهی اطلاعات بالقوه مفید از روی داده هایی است که قبلاً، ناشناخته مانده اند. این مطلب برخی از روش های فنی مانند خوشه بندی، خلاصه سازی داده ها، فراگیری قاعده های رده بندی، یافتن ارتباط شبکه ها، تحلیل تغییرات و کشف بی قاعدگی را شامل می شود » (پیاتتسکی شاپیرو، ماتئوس کریستوفر)

۲- « داده کاوی در حقیقت کشف ساختارهای جالب توجه، غیر منتظره و با ارزش از داخل مجموعه وسیعی از داده ها می باشد و فعالیتی است که اساساً با آمار و تحلیل دقیق داده ها منطبق است» هند (۱۹۹۸)

۳- « داده کاوی فرآیند کشف رابطه ها، الگوها و روندهای جدید معنی داری است که به بررسی حجم وسیعی از اطلاعات ذخیره شده در انبارهای داده با فناوری های تشخیص الگو (مانند ریاضی و آمار) می پردازد»

کشف دانش در پایگاه داده ها در جهت کشف اطلاعات مفید از مجموعه بزرگ داده هاست. دانش کشف شده می تواند قاعده ای باشد تا ویژگی های داده ها، الگوهایی که به طور متناسب رخ می دهند، خوشه بندی موضوع های درون پایگاه داده ها و غیره را توصیف می کند.

یک کاربر سیستم **KDD** بایستی درک بالایی از قلمرو داده ها به منظور انتخاب زیر مجموعه صحیحی از داده ها، رده مناسبی از الگوها و معیار خوبی برای الگوهای جالب داشته باشد. بنابراین سیستم **KDD** باید ابزارهایی با اثر تعاملی داشته باشد نه سیستم های تجزیه و تحلیل خودکار. لذا کشف دانش از پایگاه داده ها باید مثل یک فرآیند شامل گام های زیر باشد:

۱- درک قلمرو

۲- آماده کردن مجموعه داده ها

۳- کشف الگوها (داده کاوی)

۴- پردازش بعد از کشف الگو

۵- استفاده از نتایج .

نرم‌افزارهای داده‌کاوی

طی سال‌های گذشته جریان سریعی از تمایل به داده‌کاوی در بازارهای نرم‌افزاری به وجود آمده است. بیشتر کاربران نرم‌افزارهای داده‌کاو با تفکر استفاده تجاری از این نرم‌افزارها، خواهان استفاده از آن شده‌اند. نرم‌افزارهای داده‌کاو معمولاً سه روش مختلف را برای استفاده از داده‌کاوی به کار می‌برند. ۱) اکتشاف ۲) استفاده از مدل‌های پیشگویی ۳) استفاده از آنالیز بحث و جدل.

اکتشاف، فرآیند جستجو در داده‌هاست تا الگوهای مخفی موجود در داده‌ها را بدون هیچ ایدهٔ از پیش تعیین شده‌ای مشخص نماید. در نرم‌افزارهای داده‌کاوی مبتنی بر مدل‌های پیشگویی، الگوهایی که از یک بانک داده کشف می‌شوند، برای پیش‌بینی آینده به کار می‌روند. مدل‌های پیش‌بینی به کاربر اجازه می‌دهند تا داده‌های نامشخص را به کار ببرد و این مقادیر نامشخص توسط نرم‌افزار کشف شود. در مدل‌های جدلی نیز الگوهای یافت شده از داده‌ها برای تعیین مقادیر غیرعادی به کار می‌رود. برای تعیین مقادیر غیر عادی، ابتدا می‌بایست مقادیر عادی شناخته شود تا بر این اساس مقادیر غیرعادی و منحرف شناخته شوند.

نرم‌افزارهای داده‌کاو در حال حاضر از فعالیت کمتری نسبت به سایر نرم‌افزارهای هوشمند برخوردار هستند. با این وجود فعالیت تجاری این نرم‌افزار را می‌توان در شش بخش کلی، دسته‌بندی داده‌ها، برآورد مقادیر نامشخص، پیش‌بینی مقادیر نامشخص, گروه‌بندی تقریبی داده‌ها، خوشه‌بندی داده‌ها و تشریح روابط بین داده‌ها تقسیم کرد.

شرکت‌ها، سازمان‌ها، دانشگاه‌ها و مؤسسات آموزش عالی امروزی غرق در انبوه داده‌ها و اطلاعاتی هستند که استفاده از آنها در بیشتر موارد محدود به انجام کارهای جاری می‌باشد و هنوز از داده‌ها در تصمیم‌گیری استراتژیک استفاده نمی‌شود. داده‌کاوی که استفاده از آن روز به روز توسعه می‌یابد می‌تواند به استفاده از اطلاعات موجود در مؤسسات و مراکز آموزش عالی در زمینه‌های تصمیم‌گیری استراتژیک منجر شود.

عبارت داده کاوی مترادف با یکی از عبارت های استخراج دانش، برداشت اطلاعات، وارسی داده ها و حتی لایروبی کردن داده هاست که در حقیقت کشف دانش در پایگاه داده ها۱۱ (KDD) را توصیف می کند. بنابراین ایده ای که مبنای داده کاوی است یک فرآیند با اهمیت از شناخت الگوهای بالقوه مفید، تازه و درنهایت قابل درک در داده هاست. واژه کشف دانش در پایگاه داده ها در اوایل دهه ۸۰ در مراجعه به مفهوم کلی، گسترده، سطح بالا و به دنبال جستجوی دانش در اطلاعات شکل گرفته است. داده کاوی کاربرد سطح بالای فنون و ابزار بکار برده شده برای معرفی و تحلیل داده ها ی تصمیم گیرندگان است .اصطلاح داده کاوی را آمار شناسان، تحلیل گران داده ها و انجمن سیستم های اطلاعات مدیریت به کار برده اند در حالی که پژوهشگران یادگیری ماشین و هوش مصنوعی از KDD بیشتر استفاده می کنند. در ادامه چند تعریف از داده کاوی ارائه می شود.

»داده کاوی یا به تعبیر دیگر کشف دانش در پایگاه داده ها، استخراج غیر بدیهی اطلاعات بالقوه مفید از روی داده هایی است که قبلاً، ناشناخته مانده اند. این مطلب برخی از روش های فنی مانند خوشه بندی، خلاصه سازی داده ها، فراگیری قاعده های رده بندی، یافتن ارتباط شبکه ها، تحلیل تغییرات و کشف بی قاعدگی را شامل می شود « (پیاتتسکی شاپیرو، ماتئوس کریستوفر)

»داده کاوی در حقیقت کشف ساختارهای جالب توجه، غیر منتظره و با ارزش از داخل مجموعه وسیعی از داده ها می باشد و فعالیتی است که اساساً با آمار و تحلیل دقیق داده ها منطبق است» هند (۱۹۹۸)

یک کاربر سیستم KDD بایستی درک بالایی از قلمرو داده ها به منظور انتخاب زیر مجموعه صحیحی از داده ها، رده مناسبی از الگوها و معیار خوبی برای الگوهای جالب داشته باشد. بنابراین سیستم KDD باید ابزارهایی با اثر تعاملی داشته باشد نه سیستم های تجزیه و تحلیل خودکار. لذا کشف دانش از پایگاه داده ها باید مثل یک فرآیند شامل گام های زیر باشد :

درک قلمرو

آماده کردن مجموعه داده ها

کشف الگوها (داده کاوی)

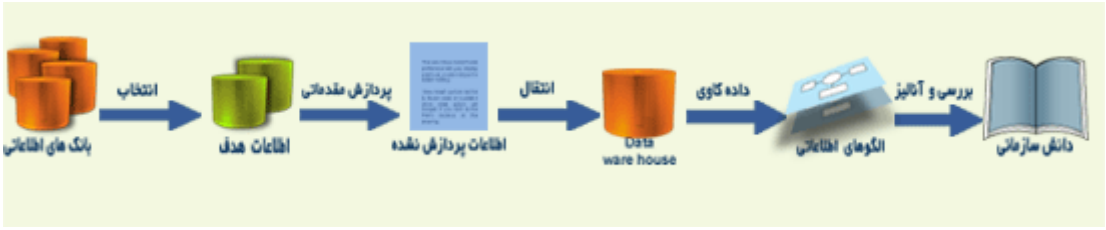
پردازش بعد از کشف الگو

استفاده از نتایج

کاوش‌های ماشینی در داده‌ها یا داده‌کاوی (Data mining) را باید یکی از سامانه‌های هوشمند (Intelligent systems) دانست. سامانه‌های هوشمند زیر شاخه‌ایست بزرگ و پرکاربرد از یادگیری ماشینی که خود زمینه‌ای‌ست در هوش مصنوعی. زمینه علمی جدید و پهناور یادگیری ماشینی (که "کاوش‌های ماشینی در داده‌ها" بخشی ست بزرگ از زیر شاخه سامانه‌های هوشمند آن ست)، به واقع همان امتداد و استمرار دانش کهن و همه جا گیر آمار است در جهت ماشینی کردن یادگیری، تعلّم، و سرانجام، دانش.

داده کاوی به عنوان مهمترین کاربرد Data Warehouse یاانباره های داده شناخته می شود . به وسیله داده کاوی داده های موجود مورد تحلیل قرار می گیرند تا روندهای احتمالی، ارتباط‌های غیر محسوس و الگو های مخفی داده ها از بین انبوه داده ها، شناسایی شوند.

در این فرایند از الگوریتم های پیچیده ریاضی و آماری استفاده می شود تا داده ها تبدیل به دانش سازمان شوند.



امروزه با حجم عظیمی از داده ها روبرو هستیم. برای استفاده از آنها به ابزارهای کشف دانش نیاز داریم. داده کاوی به عنوان یک توانایی پیشرفته در تحلیل داده و کشف دانش مورد استفاده قرار می گیرد. داده کاوی در علوم (ستاره شناسی،...)در تجارت (تبلیغات، مدیریت ارتباط با مشتری،...) در وب (موتورهای جستجو،...) در مسایل دولتی (فعالیتهای ضد تروریستی،...) کاربرد دارد. عبارت داده کاوی شباهت به استخراج زغال سنگ و طلا دارد. داده کاوی نیز اطلاعات را که در انبارهای داده مدفون شده است، استخراج می کند.

در واقع هـــــدف از داده کاوی ایجاد مدل هایی برای تصمیم گیری است. این مدلها رفتارهای آینده را براساس تحلیلهای گذشته پیش بینی می کنند. به کاربردن داده کاوی به عنوان اهرمی برای آماده سازی داده ها و تکمیل قابلیتهای انباره داده (DATA WAREHOUSE) ، بهترین موقعیت را برای به دست آوردن برتریهای رقابتی ایجاد می کند.

سیستم های بانک داده (BASE DATA) ، نقشی کلیدی در سیستم های مدیریت و انبار داده، بازی می کنند. یک سیستم بانک داده، شامل فایل های بانک داده و سیستم های مدیریت بانک داده است.

اغلب تجارت ها به تصمیم گیریهای استراتژیک و یا اتخاذ خط مشی های جدید برای خدمت رسانی بهتر به مشتریان نیاز دارند. به عنوان مثال فروشگاهها آرایش مغازه خود را برای ایجاد میل بیشتر به خرید مجدداً طراحی می کنند و یا خطوط هواپیمایی تسهیلات خاصی را برای مشتریان جهت پروازهای مکرر آنها در نظر می گیرند. این دو مثال به داده هایی در مورد رفتار مصرفی گذشته مشتریان برای تعیین الگوهایی به وسیله داده کاوی، نیاز دارد. براساس این الگوها تصمیمـــات لازم

اتخاذ می شود. در واقع ابزار داده کـــاوی، داده را می گیرد و یک تصویر از واقعیت به شکل مدل می سازد، این مدل روابط موجود در داده ها را شرح می دهد.

از نظر فرایندی فعالیتهای داده کاوی به سه طبقه بندی عمومی تقسیم می شوند:

اکتشاف :

فرایند جستجو در یک بانک داده برای یافتن الگوهای پنهان، بدون داشتن یک فرضیه از پیش تعیین شده درباره اینکه این الگو ممکن است چه باشد.

مانند تحلیلهایی که برحسب کالاهای خریداری شده صورت می گیرد، اینگونه تحلیلهای سبدی نشانگر مواردیست که مشتری تمایل به خرید آنها دارند. این اطلاعات می تواند به بهبود موجودی، استراتژی طراحی، آرایش فروشگاه و تبلیغات منجر گردد.

مدل پیش بینی :

فرایندی که الگوهای کشف شـــده از بانک داده را می گیرد و آنها را برای پیش بینی آینده به کار می برد.

مانند پیش بینی فروش در خرده فروشی، الگوهای کشف شده برای فروش به آنها کمک می کند تا تصمیماتی را در رابطه با موجودی اتخاذ کنند.

تحلیلهای دادگاهی :

به فرایند به کارگیری الگوهای استخراج شده برای یافتن عوامـــل داده ای نامعقول و متناقض مربوط می شود. مانند شناسایی و تشخیص کلاهبرداری در موسسات مالی. کلاهبرداری به میزان زیادی پرهزینه و زیان آور است، بانکها می توانند با تحلیل دادوستدهای جعلی گذشته الگوهایی را برای تشخیص و کشف کلاهبرداری به دست آورند.

از نمایی دیگر، داده کاوی ، بعنوان روشی در استخراج دانش از متون، یکی از موضوعات مهم در گستره ای از اعمال مدیریت اطلاعات است. در این میان آنچه از اهمیت فوق العاده ای برخوردار است ارایه راه‌کارهایی برای مواجه با این حجم عظیم اطلاعاتی و استفاده بهینه از اطلاعات در جهت خلق دانش، تولید سینرجی و در نهایت افزایش خرد جمعی است.

در سالهای اخیر اهمیت متون به عنوان منابع با پتانسیل اطلاعاتی بسیار بالا به نحو گسترده‌ای مورد توجه قرار گرفته به طوری که کشف دانش از متون به عنوان یکی از مهمترین فعالیتهای محققین حوزه هوش مصنوعی و فناوری اطلاعات قرار گرفته است. تحقیقات بسیاری صورت گرفته اما محدوده فعالیت بقدری گسترده است که نیازمند توجه بیشتری می‌باشد.

امروزه محققان به این مسئله معترفند که با وجود انجام تحقیقات بی وقفه در زمینه کاری خود، نمی‌توانند همزمان با پیشرفت دانش، معلومات خود را به روز نگاه دارند. بعنوان مثال بانک اطلاعاتی Medline در حال حاضر حاوی ۱۰ میلیون چکیده مقاله است و هر هفته بین هفت تا هشت هزار چکیده مقاله به این بانک اطلاعاتی افزوده می‌شود. در این بین شاید همه مقالات مربوط به یک دانش خاص نباشند، اما تعداد مقالات تخصصی که در حوزه تحقیق یک دانش خاص قرار می‌گیرد به اندازه ای است که یک نفر نمی‌تواند ادعا کند همه آنها را مطالعه کرده است بعلاوه نقش مطالعات عمیق و گسترده و استخراج ایده ها و دانش جدید از مطالب مطالعه شده بر کسی پوشیده نیست.

در این میان اینترنت بعنوان بزرگترین منبع اطلاعاتی همگانی، تشکیل یافته از صد ها میلیون صفحه اطلاعات است که به جهت همگانی بودن آن و نبود آینده‌نگری کافی در زمان تشکیل و رشد آن ، متحمل نگاهداری اطلاعات نویسندگان، محققان ، اندیشمندان و غیره به همان نحوی که آنها می نوشتند گردید. نبود یک استاندارد همه جانبه و دقیق در تنظیم متون و قرار گیری این مجموعه عظیم بصورتی غیر ساختیافته و یا بعضاً نیمه ساخت یافته، جامعه اطلاعاتی را دچار نوعی سردر گمی و مشکل

در دستیابی به اطلاعات مورد نیاز کرده بطوری‌که برای یافتن مطالب مورد نظر خود متحمل هزینه های زمانی بسیاری می‌گردند. محققان به ارایه راه کارهایی برای ساخت یافته کردن اطلاعات نمودند و با ارایه زبانهای نشانه گذاری استاندارد نظیر XML تا حد زیادی جلوی این از هم پاشیدگی اطلاعاتی را گرفتند اما آنچه همچنان باقی است وجود بسیاری از متون غیر ساخت‌یافته می‌باشد؛ در همین راستا ارایه ابزاهایی که با بررسی متون بتوانند تحلیلی روی آنها انجام دهند منجر به شکل گیری زمینه‌ای جدید در هوش مصنوعی و فناوری اطلاعات گردیده که به یادگیری متن معروف است.

این حوزه تمام فعالیتهایی که به نوعی به دنبال کسب دانش از متن هستند را شامل می‌گردد. آنالیز داده های متنی توسط تکنیکهای یادگیری ماشین، بازیابی اطلاعات هوشمند، پردازش زبان طبیعی یا روشهای مرتبط دیگر همگی در زمره مقوله یادگیری متن قرار می‌گیرند. یکی از روشهایی که ذکر گردید، استفاده از تکنیکهای یادگیری ماشین در زمینه پردازش متن است، مسئله قابل تامل این است که این تکنیکها در ابتدا در مورد داده های ساخت‌یافته به کار گرفته شدند و علمی به نام داده کاوی را بوجود آوردند. داده های ساخت‌یافته به داده هایی اطلاق می‌گردد که بطور کاملاً مستقل از همدیگر ولی یکسان از لحاظ ساختاری در یک محل گردآوری شده اند.

انواع بانکهای اطلاعاتی را می‌توان نمونه هایی از این دسته اطلاعات نام برد. در اینصورت مسئله داده کاوی عبارت از کسب اطلاعات و دانش از این مجموعه ساخت یافته. اما در مورد متون که عمدتاً غیر ساخت یافته یا نیمه ساخت یافته هستند ابتدا باید توسط روشهایی ، آنها را ساختارمند نمود و سپس از این روشها برای استخراج اطلاعات و دانش از آنها استفاده کرد. به هر حال استفاده از داده کاوی در مورد متن خود شاخه ای دیگر را در علوم هوش مصنوعی بوجود آورد به نام متن کاوی . از جمله فعالیتهای بسیار مهم در این زمینه، طبقه بندی (دسته بندی) متن می‌باشد.

طبقه بندی متن، یعنی انتساب اسناد متنی بر اساس محتوی به یک یا چند طبقه از قبل تعیین شده، یکی از مهمترین مسایل در متن کاوی است؛ مرتب کردن بلادرنگ نامه های الکترونیکی یا فایلها در سلسله مراتبی از پوشه ها، تشخیص موضوع متن، جستجوی ساختیافته و/ یا پیدا کردن اسنادی که در راستای علایق کاربر میباشد، از جمله کاربردهای مبحث طبقه بندی (دسته بندی-کلاسه بندی) متن است. در بسیاری از موارد ، افراد حرفه ای آموزش دیده، برای طبقه بندی متون جدید به کار گرفته می‌شوند. این فرآیند بسیار زمان بر و پر هزینه است و لذا کاربرد خود را محدود می‌سازد، به همین منظور علاقه روزافزونی به توسعه فناوری هایی در دسته بندی خودکار متن ابراز میشود.

در هر حال در جوامع اطلاعاتی امروزی آنچه از اهمیت روزافزونی برخوردار است، اطلاعات و تبادل آن است و در این راستا به توسعه فناوری های مرتبط پرداخته می‌شود، اما یک مرحله کاملاً جدید تر و کاملاً مورد توجه جوامع فرا صنعتی، خلق دانش جدید از اطلاعات قبلی است که این جوامع آنرا کلید موفقیت خود در آینده دانسته و به سختی در این زمینه فعالیت می‌نمایند. بر ما است تا ضمن ارتقای فناوری اطلاعات در کشور و ایجاد زیر ساختهای لازمه در اسرع وقت، به اینگونه مسائل جدی‌تر که در زمره Information High Technology قرار می گیرند، بپردازیم.

طی سال‌های گذشته جریان سریعی از تمایل به داده‌کاوی در بازارهای نرم‌افزاری به وجود آمده است. بیشتر کاربران نرم‌افزارهای داده‌کاو با تفکر استفاده تجاری از این نرم‌افزارها، خواهان استفاده از آن شده‌اند. نرم‌افزارهای داده‌کاو معمولاً سه روش مختلف را برای استفاده از داده‌کاوی به کار می‌برند. ۱) اکتشاف ۲) استفاده از مدل‌های پیشگویی ۳) استفاده از آنالیز بحث و جدل.

اکتشاف، فرآیند جستجو در داده‌هاست تا الگوهای مخفی موجود در داده‌ها را بدون هیچ ایدهٔ از پیش تعیین شده‌ای مشخص نماید. در نرم‌افزارهای داده‌کاوی مبتنی بر مدل‌های پیشگویی، الگوهایی که از یک بانک داده کشف می‌شوند، برای پیش‌بینی آینده به کار می‌روند .مدل‌های پیش‌بینی به

کاربر اجازه می‌دهند تا داده‌های نامشخص را به کار ببرد و این مقادیر نامشخص توسط نرم‌افزار کشف شود.

در مدل‌های جدلی نیز الگوهای یافت شده از داده‌ها برای تعیین مقادیر غیرعادی به کار می‌رود. برای تعیین مقادیر غیر عادی، ابتدا می‌بایست مقادیر عادی شناخته شود تا بر این اساس مقادیر غیرعادی و منحرف شناخته شوند.

نرم‌افزارهای داده‌کاو در حال حاضر از فعالیت کمتری نسبت به سایر نرم‌افزارهای هوشمند برخوردار هستند. با این وجود فعالیت تجاری این نرم‌افزار را می‌توان در شش بخش کلی، دسته‌بندی داده‌ها، برآورد مقادیر نامشخص، پیش‌بینی مقادیر نامشخص, گروه‌بندی تقریبی داده‌ها، خوشه‌بندی داده‌ها و تشریح روابط بین داده‌ها تقسیم کرد

داده کاوی و کشف دانش در پایگاه داده ها از جمله موضوع هایی هستند که همزمان با ایجاد و استفاده از پایگاه داده ها در اوایل دهه ۸۰ برای جستجوی دانش در داده ها شکل گرفت.

شاید بتوان لوول (۱۹۸۳) را اولین شخصی دانست که گزارشی در مورد داده کاوی تحت عنوان « شبیه سازی فعالیت داده کاوی » ارائه نمود. همزمان با او پژوهشگران و متخصصان علوم رایانه، آمار، هوش مصنوعی، یادگیری ماشین و . . . نیز به پژوهش در این زمینه و زمینه های مرتبط با آن پرداخته اند.

پژوهش جدی روی موضوع داده کاوی از اوایل دهه ۹۰ شروع شد. پژوهش ها و مطالعه های زیادی در این زمینه صورت گرفته، همچنین سمینارها، دوره های آموزشی و کنفرانس هایی نیز برگزار شده است. نتایج پایه های نظری داده کاوی در تعدادی از مقاله های پژوهشی آورده شده است. مثلاً سال ۱۹۹۱ پیاتتسکی و شاپیرو ۲ « استقلال آماری قاعده ها در داده کاوی» را بررسی نموده اند. سال ۱۹۹۵ هافمن و نش استفاده از داده کاوی و داده انبار۳ توسط بانک های آمریکا را بررسی نموده و بیان کردند که چگونه این سیستم ها برای بانک های آمریکا قدرت رقابت بیشتری ایجاد می کنند.

چت فیلد مشکلات ایجاد شده توسط داده کاوی را بررسی نمود و همچنین مقاله ای تحت عنوان « مدل های خطی غیر دقیق داده کاوی و استنباط آماری « ارایه نمود. هندری نیز دیدگاه اقتصاد سنجی روی داده کاوی را تهیه کرد. در این سال انجمن داده کاوی همزمان با اولین کنفرانس بین المللی «کشف دانش و داده کاوی» شروع به کار کرد. این کنفرانس توسعه یافته چهار دوره آموزشی بین المللی در پایگاه های داده درسال ۱۹۸۹ تا ۱۹۹۴ بود. انجمن مذکور، یک سازمان علمی به نام ACM-SIGKDDرا ایجاد نمود. سال ۱۹۹۶ ایمیلنسکی۴ و منیلا۵ دیدگاهی از داده کاوی به عنوان«پرس و جو کننده از پایگاه های استنتاجی۶» را پیشنهاد کردند. فایاد، پیاتتسکی – شاپیرو، اودورأسامی پیشرفت های کشف دانش و داده کاوی را عنوان کردند. در سال ۱۹۹۷ منیلا خلاصه ای از مطالعه روی اساس داده کاوی ارایه نمود. باربارا و همکاران نیز دیدگاه کاهش داده ها روی داده کاوی را در گزارش کاهش داده های نیوجرسی ارایه نمودند. همچنین می توان برای کاربرد داده کاوی در مدیریت مالی می توان، تحلیل داده های مالی و مدل سازی مالی بنینگاه و چاچ کز و هیگینز ۷ را ملاحظه کرد فریدمن نیز مقاله ای در ارتباط با مفهوم آمار و داده کاوی ارایه نمود. سال ۱۹۹۸ هند ۸ مقاله ای تحت عنوان « داده کاوی : آمار یا بیشتر؟ » ارائه نمود .کلینبرگ۹ پائودیمیتریو و راغان ۱۰ دیدگاه اقتصاد سنجی روی داده کاوی و عملکرد داده کاوی به عنوان یک مسئله بهینه را ارایه نمودند. در این سال نیز کنفرانس های ناحیه ای و بین المللی در مورد داده کاوی برگزار شد که از جمله می توان به کنفرانس آسیا و اقیانوسیه درباره کشف دانش و داده کاوی اشاره کرد. سال ۲۰۰۰ هند و همکاران و اسمیت بحث های مقایسه ای بین آمار و داده کاوی را ارایه کردند. سری و استاوا، کولی، رش پاند و تن استفاده از وب در کاوش داده ها و کاربردهای آن را ارایه کردند. سال ۲۰۰۲ کلادیو کانورسانو و همکاران « مدل آمیخته چندگانه جمع پذیر تعمیم یافته » برای داده کاوی را بررسی نمودند. پائلو و گیانلوکاپاسرون، « داده کاوی ساختارهای پیوند برای مدل رفتار مصرف کننده » را ارایه نمودند

ابزارهای تجاری داده کاوی **DM Commercial Tools**

در مورد ابزارهای موجود برای داده کاوی باید به این نکات توجه داشت که:

- مدل/معماری مشترکی بین آنها موجود نمی باشد.
- به منابع داده گوناگون و نه لزوماً همه گونه منبعی دسترسی دارند.
- از یک یا بیشتر الگوریتم DM پشتیبانی مینماید.
- ممکن است از تمام انواع داده پشتیبانی کند یا نکند.
- قابلیتهای مختلف اما نه تمام آنها را پشتیبانی مینماید.
- وابسته به بستر کاری هر کاربردی ممکن است با یک ابزار کار کند و با ابزار دیگر کار نکند.

ابزارهای تجاری برای داده کاوی را می توان به صورت زیر لیست کرد

- Darwin (Oracle Corp(.
- MineSet (Silicon Graphics Inc. - SGI(
- Intelligent Miner (IBM Corp(
- Enterprise Miner (SAS Institute Inc(.
- Clementine (SPSS Inc – Integral Solutions(
- DMMiner (DBMiner Technology Inc(.
- BrainMaker) California Scientific Software(
- CART (Salford Systems(
- MARS (Salford Systems(

- Scenario (Cognos Inc(.
- Web Analyst (Megaputer Intelligence Inc(.
- SurfAid Analysis (IBM corp(
- Visualizer Workstation (Computer Science Innovations, Inc(

منابع اطلاعاتی مورد استفاده

منابع اطلاعاتی گوناگونی را میتوان در زمینه داده کاوی بکار برد که عبارتند از:

- پایگاه داده های رابطه ای
- انبارهای داده
- فایلها
- وب
- پایگاه های داده شیء گرا
- چند رسانه ای

انبار داده

بسیاری از سازمانها داده های خود را از مخازن داده همگن و ناهمگن در یک مجموعه داده عمومی به نام انبار داده جمع آوری و ذخیره مینمایند.(Warehouse Data)

انبار داده شامل داده های فعلی و قبلی است که برای برنامه ریزی و پیش بینی در سیستمهای پشتیبان تصمیم گیری (Decision Support System) استفاده خواهد گردید.

پایگاه های داده سنتی پایگاه هایی عملیاتی هستند که داده های روزانه را در خود ذخیره مینمایند.

star -schema, Snow-Flakes و Galaxy مدلهای رایج در انبارهای داده هستند.

برای افزایش کارایی در DW تکنیکهای مختلفی مانند خلاصه کردن و denormalization استفاده میگردد.

پیشرفت در تکنولوژیهای داده پردازی

سازمانهای بزرگ و چند- مکانه مثل بانکها، دفاتر هواپیمایی و فروشگاههای زنجیره ای با حجم زیادی از داده ها که ناشی از عملکرد روزانه آنهاست روبرو هستند. بطور سنتی چنین داده هایی به دو دسته تقسیم شده اند:

۱. رکوردهای اصلی

۲. رکوردهای عملیاتی

فرض بر این است که رکوردهای اصلی حاوی اطلاعات پایه هستند که معمولا چندان تغییر نمی کنند در حالیکه رکوردهای عملیاتی با توجه به طبیعت عملیات تجاری حتی بطور ساعتی تغییر خواهند کرد. سیستمهای مدیریت پایگاه داده مناسب برای پیوند دادن این دو مجموعه اطلاعاتی و تهیه گزارشهای استاندارد جهت کنترل فعالیتها گسترش یافتند. سیستم اطلاعات مدیریت رایج برای پشتیبانی عملیات و سرویس دهی به چند کاربر در سطوح مختلف سازمان مبتنی بر این نظریه است.

بمنظور کمک به تصمیم گیری راهبردی، نظریه تاسیس بانک اطلاعات رکوردهای اصلی به نظریه سازماندهی دیتا مارت و انبار داده ها تغییر یافت. استخراج اطلاعات از رکوردهای عملیاتی یا پایگاههای اطلاعات عملیاتی و سازماندهی آن برای تحلیل استاندارد یا زمانی فلسفه اولیه و اصولی چنین پیشرفتهایی است. گرچه، دیتا مارت و انبار داده ها از نظر هدف و ساختار با هم متفاوتند.

دیتامارت

دیتا مارت اغلب کوچک است و بر یک موضوع یا دپارتمان خاص متمرکز است. بنابراین پاسخگوی یک نیاز داخلی است. طرح بانک اطلاعات برای یک دیتامارت حول ساختار اتصال ستاره ای ساخته شده است که بهینه برای نیازهای کاربران دپارتمان است. دیتامارت معمولا با ابزارهای کامپیوتری که انعطاف پذیری تحلیل را تامین میکنند اما ممکن است برای سازماندهی حجم بالای داده ها مناسب نباشند؛ نیرومند میشود. رکوردهای ذخیره شده در دیتامارتها بخوبی نمایه شده اند. یک دیتامارت در صورتیکه داده ها را از منابع داده ای بسیار سازماندهی شده مثل انبار داده ها بگیرد؛ دیتامارت وابسته نامیده میشود. مسلما دیتامارتهای وابسته از لحاظ ساختاری و معماری منطقی هستند. منبع دیتامارتهای وابسته تکنولوژِی بانک اطلاعات دپارتمانی است. دیتامارتهای مستقل ثابت نیستندو از لحاظ معماری بسیار با هم متفاوتند. این مساله هنگام یکپارچه سازی دیتامارتهای مستقل، مشکل ایجاد میکند. بنابراین با یکپارچه سازی ساده دیتامارتها یک انبار داده ایجاد نخواهد شد.

دیتامارت اساسا برای اهداف تاکتیکی طراحی شده است و هدفش تامین یک نیازتجاری فوری است.

انبار داده ها

یک انبار داده کاملا " متفاوت از دیتامارت است. سازماندهی انبارهای داده بگونه ایست که کلیه موضوعات حول فعالیتهای کاری سازمان را می پوشاند. انبار داده نمایانگر یک تسهیلات مرکزی است. برخلاف دیتامارت که در آن داده ها به شکل خلاصه تر و متراکم تر وجود دارند، یک انبار داده ، داده ها را در یک سطح نامتراکم ذخیره می کند.

ساختار داده ها در یک انبار داده یک ساختار لزوما" هنجار شده است. بدین معنی که ساختار و محتوای داده ها در انبار داده منعکس کننده ویژگیهای دپارتمانهای عضو نیست. داده ها در انبار داده از

نظر حجم و شکل کاملا" متفاوت از داده ها در دیتامارت هستند. دیتامارت ممکن است شامل حجم زیادی از داده های قدیمی و گذشته نگر باشد. داده ها در انبار داده اغلب بصورت نسبتا" سبک نمایه میشوند. (به بیان دیگر در عمق کمتر).

انبار داده برای اهداف برنامه ریزی بلندمدت و راهبردی طراحی میشوند. در نتیجه انبار داده برخلاف سیستم عملیات که کاربرمدار است متمرکز بر اقلام است. ساختار یک انبارداده مشخصات زیر را نشان می دهد:

- **وابستگی به زمان:**

رکوردها بر اساس یک برچسب زمانی نگهداری میشوند. وابستگی زمانی حاصل در ایجاد صفحات زمانی مفید است که درک ترتیب زمانی وقایع را تسهیل میکند.

- **غیر فرار بودن:**

رکوردهای داده در انبار داده ها هرگز بطور مستقیم روزآمد نمیشوند. برای هر تغییری در ابتدا داده های عملیاتی روزآمد میشوند و سپس بگونه ای مقتضی به انبار داده منتقل میشوند. این مساله ثبات داده ها را برای استفاده های وسیعتر تضمین میکند.

- **تمرکز موضوعی:**

داده ها از بانکهای اطلاعاتی عملیاتی بصورت گزینشی به انبار داده منتقل میشوند. این استراتژی به ایجاد یک انبار داده بر اساس یک مطلب یا موضوع خاص کمک میکند و بنابراین کاوش انبار داده ها برای پرس و جوهای موضوعی با سرعت بیشتری انجام میشود.

- **یکپارچگی:**

داده ها بگونه ای کامل سازماندهی شده اند تا با حذف موارد تکراری و چند عنوانه یکپارچگی رکوردها حفظ شود ؛ به ایجاد ارجاع های متقابل کارآمد بین رکوردها کمک نموده و ارجاع دهی را تسهیل نماید.

واضح است که انبار داده اساسا" برای پرس و جوهای پشتیبان تصمیم گیری ساخته شده است. بر این اساس سازماندهی وعملیات انبار داده چنان طراحی شده اند تا نیازهای اطلاعاتی روزمره یا معمولی را پاسخگو باشند. بدلیل حجم بسیار بالای چنین پایگاه اطلاعاتی یک سیستم کامپیوتری پیشرفته برای عملیات انبارسازی داده ها لازم است. همچنین یک بانک اطلاعات مجزا شامل ابرداده که مشخصه هایی نظیر نوع، فرمت، مکان و پدیدآورندگان داده های ذخیره شده در یک انبار داده ها را توصیف میکند نیز برای کمک به کاربران و مدیران داده ها ساخته میشود.

مشخص شد که انبار داده بدلیل اندازه و تنوعش، اگر مبتکرانه پردازش شود میتواند به تولید اطلاعاتی منجر شود که در وهله اول آشکار نیستند. با انتخاب متناسب داده ها، بکار گرفتن فنون مختلف غربال کردن و تفسیر زمینه ای ، داده ذخیره شده میتوانست منجر به کشف الگوها یا رابطه هایی شود که بینش نویی به تصمیم گیرنده دهد. این مساله نظریه توسعه عملیات داده کاوی را به موازات معدن کاوی بروز داد. ذکر این نکته لازم است که داده کاوی در اصل لزوما" نیاز به سازماندهی یک انبار داده ندارد.

عناصر داده کاوی

توصیف و کمک به پیش بینی دو کارکرد اصلی داده کاوی هستند. تحلیل داده مربوط به مشخصه های انتخابی متغیرها؛ از گذاشته و حال، و درک الگو مثالی از تحلیل توصیفی است. برآورد ارزش آینده یک متغیر و طرح ریزی کردن روند مثالی از توانایی پیشگویانه داده کاوی است. برای عملی شدن هریک از دو کارکرد فوق الذکر داده کاوی، چند گام ابتدایی اما مهم باید اجرا شوند که از این قرارند:

۱. انتخاب داده ها

۲. پاک سازی داد ها

۳. غنی سازی داده ها

۴. کد گذاری داده ها

با دارا بودن هدف کلی در مطالعه، انتخاب مجموعه داده های اصلی برای تحلیل، اولین ضرورت است. رکوردهای لازم میتواند از انبار داده ها و یا بانک اطلاعاتی عملیاتی استخراج شود. این رکوردهای داده جمع آوری شده؛ اغلب از آنچه آلودگی داده ها نامگذاری شده است رنج می برند و بنابراین لازم است پاکسازی شوند تا از یکدستی فرمت (شکلی) آنها اطمینان حاصل شود، موارد تکراری حذف شده و کنترل سازگاری دامنه بعمل آید. ممکن است داده های گردآوری شده از جنبه های خاصی ناقص یا ناکافی باشند. در این صورت داده های مشخصی باید گردآوری شوند تا بانک اطلاعات اصلی را تکمیل کنند. منابع مناسب برای این منظور باید شناسایی شوند. این فرایند مرحله غنی سازی داده ها را تکمیل میکند. یک سیستم کدگذاری مناسب معمولا" جهت انتقال داده ها به فرم ساختار-بندی شده جدید؛ متناسب برای عملیات داده کاوی تعبیه میشود .

فنون داده کاوی

ممکن است متوجه شده باشید که فنون داده کاوی یک گروه نامتجانس را شکل میدهند چرا که هر تکنیکی که بتواند بینش جدیدی از داده ها را استخراج کند میتواند داده کاوی به حساب آید. برخی از ابزارهای رایج بکار گرفته شده تحت عنوان داده کاوی عبارتند از:

ابزارهای پرس و جو:

ابزارهای متداول زبان پرس و جوی ساختاربندی شده در ابتدا برای انجام تحلیلهای اولیه بکار گرفته شدند که می تواند مسیرهایی برای تفحص بیشتر نشان دهد.

فنون آماری:

مشخصات اصلی داده ها لازمست با کاربرد انواع مختلفی از تحلیلهای آماری شامل جدول بندی ساده و متقاطع داده ها و محاسبه پارامترهای آماری مهم بدست آید.

مصور سازی:

با نمایش داده ها در قالب نمودارها و عکسها مانند نمودار پراکندگی؛ گروه بندی داده ها در خوشه های متناسب تسهیل میشود. استنباط عمیق تر ممکن است با بکارگیری تکنیکهای گرافیکی پیشرفته حاصل شود.

پردازش تحلیلی پیوسته:

از آنجا که مجموعه داده ها ممکن است روابط چندین بعدی داشته باشند، روشهای متعددی برای ترکیب کردن آنها وجود دارد. ابزارهای پردازش تحلیلی پیوسته به ذخیره چنین ترکیباتی کمک میکند و ابزارهای ابتدا-انتهای پیوسته برای انجام پرس و جو ایجاد میکند. اما این ابزارها هیچ دانش جدیدی ایجاد نمی کنند.

یادگیری مبتنی بر مورد:

این تکنیک مشخصات گروههای داده ها را تحلیل میکند و به پیش بینی هر نهاد واقع شده در همسایگی شان کمک میکند. الگوریتمهایی که استراتژی یادگیری تعاملی را برای کاوش در یک فضای چندین بعدی بکار میگیرند برای این منظور مفیدند.

درختان تصمیم گیری:

این تکنیک بخشهای مختلف فهرست پاسخهای موفق داده شده مربوط به یک پرس و جو را بازیابی می کند و به این ترتیب به ارزیابی صحیح گزینه های مختلف کمک میکند.

قوانین وابستگی:

اغلب مشاهده میشود که یک وابستگی نزدیک (مثبت یا منفی) بین مجموعه ای از داده های معین وجود دارد. بنابراین قوانین رسمی وابستگی برای تولید الگوهای جدید ساخته و بکار گرفته میشوند.

شبکه های عصبی :

این یک الگوریتم یادگیری ماشینی است که عملکرد خودش را بر اساس کاربرد و ارزیابی نتایج بهبود می بخشد.

الگوریتم ژنتیکی:

این هم تکنیک مفید دیگری برای پیش بینی هدف است. به این ترتیب که با یک گروه یا خوشه شروع میشود و رشدش در آینده را با حضور در برخی مراحل فرایند محاسبه احتمال جهش تصادفی؛ همانطور که در تکامل طبیعی فرض میشود طرح ریزی می نماید. این تکنیک به چند روش میتواند عملی شود. و ترکیب غیرقابل انتظار یا نادری را از عواملی که در حال وقوع بوده و مسیر منحنی طراحی داده ها را تغییر میدهند؛ منعکس میکند.

گام نهایی فرایند داده کاوی، گزارش دادن است. گزارش شامل تحلیل نتایج و کاربردهای پروژه، درصورت بکارگیری آنها، است . و متن مناسب، جداول و گرافیکها را در خود جای می دهد. بیشتر اوقات گزارش دهی یک فرایند تعاملی است که تصمیم گیرنده با داده ها در پایانه کامپیوتری بازی میکند و فرم چاپی برخی نتایج واسطه محتمل را برای عملیات فوری بدست می آورد. داده کاوی در تولید چهار نوع دانش ذیل مفید است:

- دانش سطحی: کاربردهای SQL
- دانش چند وجهی: کاربردهای OALP
- دانش نهان : تشخیص الگو و کاربردهای الگوریتم یادگیری ماشینی
- دانش عمیق :کاربردهای الگوریتم بهینه سازی داخلی

از آنجا که داده کاوی با بانکهای اطلاعاتی بزرگ سروکار دارد، به گونه ای ایده ال با تکنولوژی خدمت گیر–خدمت گر بکار میرود. کاربردهای عمومی داده کاوی بیشتر شامل تقسیم کردن داده ها در خوشه های مقتضی، کدگذاریهای مناسب، کاوش برای الگوها و طراحی کردن با استفاده از فنون آماری و الگوریتمهای ژنتیکی است. تعداد زیادی از بسته های نرم افزاری واجد این جنبه های ابزارهای داده کاوی با درجات متفاوتی از جامعیت در دسترس هستند. برای مثال بسته های نرم افزاری که منحصرا" برای کاربردهای OLAP در دسترس هستند عبارتند از:

Clever Path OLAP ,Oracle OLAP, DB2 OLAP Server

نرم افزارهای آماری عمومی مثل SPSS, SAS, STATISTICA با امکاناتی برای داده کاوی و بسته های نرم افزاری اختصاصی داده کاوی مثل

Miner3, Text Mining Software, Enterprise Data Mining software, Weka, Insightful PolyAnalyst 4.6

مفید هستند.

محدودیت های داده کاوی

کاربرد داده کاوی با چند عامل محدود شده است. اولین مورد به سخت افزار و نرم افزار لازم و موقعیت بانک اطلاعاتی مربوط میشود . برای مثال در هند، داده های غیر مجتمع که برای کاربردهای داده کاوی لازم است ممکن است به فرم دیجیتالی در دسترس نباشد. در دسترس بودن نیروی انسانی ماهر در داده کاوی نیز مسأله مهم دیگری است. محرمانه بودن رکوردهای مراجعان ممکن است در نتیجه پردازش داده های مبتنی بر داده کاوی آسیب پذیر شود. کتابداران و مؤسسات آموزشی باید این مسأله را در نظر داشته باشند؛ چرا که در غیر اینصورت ممکن است گرفتار شکایات قانونی گردند.
محدودیت دیگراز ضعف ذاتی نهفته در ابزارهای نظری ناشی میگردد. ابزارهایی مانند یادگیری ماشینی و الگوریتمهای ژنتیکی بکار گرفته شده در فعالیتهای داده کاوی به مفاهیم وفنون منطق و آمار بستگی دارد. در این حد نتایج به روش مکانیکی تولید شده و بنابراین به یک بررسی دقیق نیاز دارند. اعتبار الگوهای بدست آمده به این طریق؛ باید آزمایش شود. چرا که که در بسیاری موارد روابط علل و معلول مشتق شده؛ از برخی استدلالات غلط ذیل رنج میبرند.

حفاظت از حریم شخصی در سیستم‌های داده‌کاوی

داده‌کاوی با استخراج موفقیت‌آمیز اطلاعات، دانش مورد نیاز برای استفاده در زمینه‌های مختلف از جمله، بازاریابی، هواشناسی، تحلیل‌های پزشکی و امنیت ملی را فراهم می‌سازد، ولی هنوز هیچ تضمینی ارایه نشده است که بتوان داده‌های خاصی را مورد داده‌کاوی قرار داد؛ بدون آن‌که به حریم خصوصی مالک آن اطلاعات تجاوز کرد. برای مثال، در یک سیستم پزشکی، نحوه انجام داده‌کاوی در اطلاعات خصوصی بیماران بدون افشای آن اطلاعات، یکی از مسائلی است که با آن روبه‌رو

هستیم. ارگان‌هایی نظیر سازمان بیمه سلامتی و بررسی وضع بهداشت در ایالات‌متحده (HIPPA) و سازمان مدیریت داده و سیستم‌های تحلیلی در اتحادیه اروپا، با درک حساسیت‌های به وجود آمده در این زمینه، مجموعه‌ای از قوانین اجباری را در زمینه مدیریت داده و تحلیل سیستم‌ها پدید آورده‌اند. این نوع نگرانی‌ها، به موازات گسترش استفاده از سیستم‌های تحلیل داده افزایش می‌یابند. سیستم‌های جمع‌آوری داده به صورت آنلاین، نمونه‌ای از ده‌ها برنامه جدیدی هستند که حریم شخصی افراد را تهدید می‌کنند. شرکت‌های معتبر از چندی پیش با به اشتراک گذاشتن روش‌ها و مدل‌های موجود برای داده‌کاوی، به دنبال کسب داده بیشتر در مورد مشتریان مشترک هستند تا بتوانند در مورد عادت‌های آن‌ها در زمینه خرید کالا اطلاعات دقیق‌تری داشته باشند .قبل از آن‌که تکنیک‌های داده‌کاوی همه‌گیر شود و کلاف سردرگم حریم شخصی افراد را تهدید کند، باید بتوان راهی برای حفاظت از حریم و اطلاعات شخصی افراد پیدا کرد .مشکل اصلی از آنجا ناشی می‌شود که چگونه می‌توان هم حریم شخصی افراد را در نظر گرفت و هم از نتایج مفید سیستم‌های داده‌کاوی بهره برد. برای برطرف کردن موانع موجود در این زمینه، تحقیقات زیادی در حال انجام است، اما در عمل سیستم‌های داده‌کاوی که بتوانند در عین حال حریم شخصی افراد را نیز حفظ کنند، هنوز در مرحله ابتدایی و آزمایشی هستند. بیشتر این تکنیک‌ها در لایه زیرین به جای بررسی مشکلات سیستم‌ها، روی ابزارهای محاسباتی و الگوریتم‌ها متمرکز شده‌اند. هدف ما از بررسی حریم شخصی، به دست آوردن یک دید سیستماتیک از نیازهای ساختاری و طراحی اصول و بررسی راه‌حل‌هایی است که بتوانند در سیستم‌های داده‌کاوی به‌طور عملی از حریم شخصی افراد محافظت کنند.

برای مثال، در یک سیستم بیمارستانی، اداره حسابداری باید فقط بتواند به داده‌های مالی بیماران دسترسی داشته باشد و به هیچ عنوان نباید به رکوردهای ثبت شده در مورد سوابق پزشکی آن‌ها دسترسی داشته باشد. توسعه و ایجاد قوانین مؤثر برای دسترسی درست سرورهای داده‌کاوی به

داده‌های انبارهای داده، یکی از مشکلاتی است که تحقیق در مورد آن به صورت ارسال و دریافت گسترده پیشنهادها، در حال انجام است .

به علاوه، یک سرور داده‌کاوی ممکن است با ایجاد مدل‌های داده‌کاوی روی سرور انبار داده، داده‌های آن را با سرورهای داده‌کاوی دیگر در سیستم‌های دیگر به اشتراک بگذارد. انگیزه اصلی از به اشتراک گذاشتن داده در این مدل‌ها، ایجاد مدل‌های مشابه برای داده‌کاوی در بین سیستم‌ها است .

برای مثال، شرکت‌های اجاره‌دهنده سرور، ممکن است بخواهند روش‌های داده‌کاوی خود روی رکوردهای مشتریان را به اشتراک بگذارند تا به این ترتیب یک مدل جهانی داده‌کاوی در مورد رفتار مشتریان ایجاد کنند که به نفع همه شرکت‌ها خواهد بود. همان‌طورکه شکل ۱ نشان می‌دهد، به اشتراک گذاشتن داده در بالاترین لایه رخ می‌دهد که در آن هر سرور داده‌کاوی از مدل داده‌کاوی مخصوص خود استفاده می‌کند. بنابراین در اینجا «به اشتراک گذاشتن» به معنی به اشتراک گذاشتن مدل‌های داده‌کاوی محلی است، نه به اشتراک گذاشتن داده‌های خام .

فصل دوم : کاربردهای داده کاوی

داده کاوی کاربردهای مختلفی دارد که اهم کاربردهای آن:

۱- کشف تقلب (کلاه برداری) و آنالیز ریسک

کشف تقلب کارتهای اعتباری

کشف پولشویی

ریسک پرداخت وام

۲- خــرده فروشی (تک فروشی)

فروش و تبلیغ

کوپن

۳-آنالیز بازار استوک

۴- تشخیص جرائم .

۵- پیش بینی سیل.

۶- ارتباطات راه دور

۷- تشخیص طبی و درمان.

۸- آنالیز داده DNA و زیست پزشکی(Biomedical).

چه ژنهایی با ژنهای دیگر همزمان رخ میدهند.

ترتیب عملیات ژنتیکی در مراحل بیماری چیست.

۹- وب کاوی Web Mining

ارتباط بین صفحات گوناگون چیست.

مشخصات صفحه وب چیست.

توزیع اطلاعات در وب چگونه است.

برای آشنایی بیشتر با داده کاوی چند کاربرد مهم و کاربردی آن را مورد مطالعه قرار می دهیم:

کاربرد داده کاوی در کسب و کار هوشمند بانک

با رشد فزاینده حجم داده‌ها در سیستمهای متنوع کسب و کار، و همچنین نیاز روز افزون جهت دستیابی به اطلاعات ارزشمند و معرفت از این داده‌های خام، داده کاوی به عنوان روشی مهم و پرکاربرد برای استخراج اطلاعات و ارضاء این نیاز مطرح شده است. در واقع داده کاوی (Data Mining)بخشی از فرایند استخراج معرفت (Knowledge Discovery)است که در آن الگوهای مفید و ضمنی در پایگاه داده ها جستجو می‌شوند .در این میان با افزایش کاربرد سیستمهای اطلاعات جغرافیایی، پایگاه‌های بزرگی از داده‌های متنوع جغرافیایی در دسترس قرارگرفته‌اند که کمک شایانی به انجام تحلیل‌های کامل‌تر و دقیق‌تر می‌نمایند.داده کاوی روی داده‌هایی که دارای یک یا چند ویژگی مکانی، فضایی و یا جغرافیایی باشند، داده کاوی فضایی (Spatial Data Mining)نامیده می‌شود و خروجی آن اطلاعات و معرفتی است که دارای خصوصیات فضایی و جغرافیایی، مانند مکان، جهت، فاصله، شکل هندسی و مانند آن می باشد.

برای مثال فرض کنید به دنبال بررسی و اجرای یک روش داده کاوی پیشرفته روی داده‌های فضایی موجود در بانک ملت ایران می‌باشد که با داده‌های مختلف بانکی از قبیل مکان شعب، شاخصهای بانکی مانند درآمد، سود، هزینه، تعداد کارکنان، میزان مراجعه و مانند آن تلفیق خواهند شد.

بدین معنی که بعد از انجام مراحل لازم جهت آماده سازی داده‌ها -با ملاحضات لازم به دلیل فضایی بودن آنها- برای عملیات داده کاوی، شامل پردازش و پاکسازی داده‌ها (Data Processing and Cleaning)و ساخت انبار داده‌ها(Data warehousing)، و همچنین در نظر گرفتن روشهای دسترسی به داده‌های فضایی (Spatial Data Access)، الگوریتمی برای استخراج قوانین وابستگی (Association Rule Mining)توسعه و پیاده سازی خواهد شد و از آن برای کشف روابط موجود ما بین مقادیر مختلف فضایی و جغرافیایی مانند ترکیب جمعیتی، کاربری‌های منطقه، وضعیت سنی، درآمد، تحصیلات، موقعیت رقبا، شبکه معابر و مانند آن از یک طرف و شاخصهای بانکی شعب مانند سود، هزینه، درآمد، کارایی و مانند آن از طرفی دیگر استفاده خواهد شد. دانش استخراج شده از این فرایند، در تصمیم گیریی های مختلف مدیران در حوزه مدیریت شعب، مانند مکانیابی، توسعه، تلفیق و تنظیم شعب، کاربرد و اهمیت بالایی خواهد داشت.

داده کاوی درمدیریت ارتباط بامشتری

داده کاوی یکی از عناصر مدیریت ارتباط با مشتری است و می تواند به حرکت شرکتها به سمت مشتری محوری کمک کند.

داده های خام از منابع مختلفی جمع آوری می شوند و از طریق استخراج، ترجمه و فرایندهای فراخوانی به انبار داده این مدیریت وارد می شوند. در بخش مهیـــاسازی داده، داده ها از انبار خارج شده و به صورت یک فرمت مناسب برای داده کاوی در می آیند.

بخش کشف الگو شامل چهار لایه است:

۱ - سوالهای تجاری مانند توصیف مشتری

۲ - کاربردها مانند امتیازدهی، پیش گویی

۳ – روشها مانند سری های زمانی، طبقه بندی

۴ – الگوریتم ها.

در این بخش روشهای داده کاوی با کاربرد مخصوص خود برای پاسخ به سوالهای تجاری که به ذهن می رسند، الگوریتم هایی را استخراج می کنند و از این الگوریتم ها برای ساخت الگو استفاده می شود. در بخش تجزیه و تحلیل الگو، الگوها به یک دانش مفید و قابل استفاده تبدیل می شوند و پس از بهبود آنها، الگوهایی که کارا محسوب می شوند در یک سیستم اجرایی به کار گرفته خواهند شد.

رابطه مشتری با زمان تغییر می کند و چنانچه تجارت و مشتری درباره یکدیگر بیشتر بدانند این رابطه تکامل و رشد می یابد. چرخه زندگی مشتری چارچوب خوبی برای به کارگیری داده کاوی در مدیریت ارتباط با مشتری فراهم می کند. در بخش ورودی داده کاوی، چرخه زندگی مشتری می گوید چه اطلاعاتی در دسترس است و در بخش خروجی آن، چرخه زندگی می گوید چه چیزی احتمالاً جالب توجه است و چه تصمیماتی باید گرفته شود. داده کاوی می تواند سودآوری مشتری های بالقوه را که می توانند به مشتریان بالفعل تبدیل شوند، پیش بینی کند و اینکه تا چه مدت به صورت مشتریان وفادار خواهند ماند و چگونه احتمالاً ما را ترک خواهند کرد. بعضی از مشتریان مرتباً مراجعاتشان را به شرکتها برای کسب مزیتهایی که طی رقابت میان آنها به وجود می آید، تغییر می دهند. در این صورت شرکتها می توانند هدفشان را روی مشتریانی متمرکز کنند که سودآوری بیشتری دارند.

بنابراین می توان از طریق داده کاوی ارزش مشتریان را تعیین، رفتار آینده آنها را پیش بینی و تصمیمات آگاهانه ای را در این رابطه اتخاذ کرد.

کاربردهای داده کاوی در کتابخانه ها و محیط های دانشگاهی

داده کاوی در ابتدا از حوزه تجارت برخاست اما کاربردهای آن در سایر حوزه هائی که به گردآوری حجم وسیعی از داده هائی می پردازند که دستخوش تغییرات پویا نیز می گردند؛ مفید شناخته شد. بخشهایی مثل بانکداری، تجارت الکترونیک، تجارت سهام، بیمارستان و هتل از این نمونه اند. انتظار میرود که استفاده از داده کاوی در بخش آموزش بطور عام امکانهای جدید بسیاری ارائه دهد. برخی کاربردهای داده کاوی در کتابخانه ها و قسمت اداری آموزش در ذیل مورد بحث قرار گرفته اند.

عملیات کتابداری بطور کلی شامل مدیریت مدارک، ارائه خدمات و امور اداره و نگهداری است. هر کدام از این کارکردها با انواع مختلفی از داده ها سروکار دارد و بطور جداگانه پردازش میشود. اگرچه، انجام تحلیل ترکیبی براین مجموعه های داده نیز میتواند افق تازه ای را بگشاید که به طرح خدمات جدید و تحول رویه ها و عملیات جاری کمک نماید. جدول یک برخی از کاربردهای ممکن داده کاوی را که میتواند در کتابداری مفید باشد ارائه میکند.

کاربردهای داده کاوی در کتابخانه ها

بانک اطلاعاتی	کاربرد متصور
گردآوری منابع	برای تعیین نقاط قوت و ضعف مجموعه
استفاده از مجموعه	برای ایجاد رابطه بین خواننده، منابع کتابخانه و زمان مشخصی از سال
امانت بین کتابخانه ای	برای تحلیل سفارشهای پاسخ داده شده و سفارشهای دریافت شده
داده های بخش امانت	برای پیش بینی روند بازگشت منابع
داده های هزینه	برای نشان دادن منابع مالی بکار گرفته شده

داده کاوی میتواند برای پاسخ دادن به یک سوال خاص مربوط به کتابخانه و نیز برای کشف روندهای عمومی که به تصمیم گیری کمک میکنند، استفاده شود . برای مثال سوال می تواند چنین باشد:

امکان اینکه امانت گیرندگان منابع را یک هفته بعد از تاریخ عودت برگردانند تا نامه های یادآوری کمتری فرستاده شود چقدر است؟

یا میزان اشتراک مورد انتظار برای نشریات بین المللی انتخاب شده برای سال آینده چقدر است؟ درک الگوی استفاده کلی مجلات الکترونیکی یا تحلیل درخواستهای اعضا برای میکرو فیلمها طی ۵ سال گذشته نیز همگی مثالهایی از کشف روندهای عمومی اند.

دامنه تحلیل استنادی هم میتواند با استفاده از داده کاوی گسترش داده شود. در ارتباط با کتابخانه ها، وب کاوی حوزه دیگری از علاقمندی است. وب کاوی شامل محتوا کاوی وب، ساختار کاوی وب و استفاده کاوی وب با توجه به یک موضوع خاص است که در طراحی خدمات جدید مبتنی بر وب کمک خواهد کرد.

داده کاوی و مدیریت موسسات دانشگاهی

اداره موسسات دانشگاهی کار پیچیده ای است. در این موسسات دائما" نیاز به درآمدزایی و خود-کارآمدی و کاهش وابستگی به بودجه دولتی احساس میشود. این مساله کنترل دائمی جنبه های مختلف هر فعالیت و پروژه را می طلبد. بانکهای اطلاعاتی برای چنین موسساتی مربوط به دانشجویان، دانشکده، اساتید و کارمندان، تعداد رشته ها و چند مورد دیگر است . ارزیابی تقاضا و وضعیت عرضه نقش مهمی بازی میکند. مرور بانکهای اطلاعاتی نمونه در جدول ۲ نمایانگر کاربردهای بالقوه داده کاویست.

کاربردهای داده کاوی در موسسات دانشگاهی

بانک اطلاعاتی	کاربرد متصور
ثبت نام دانشگاهی	برای درک رابطه های جمعیت شناختی، اقتصادی و اجتماعی
کارایی دانشگاهی	برای ایجاد رابطه بین عوامل اقتصادی-اجتماعی و نمرات اخذ شده
بانک سوالات	برای تعیین میزان مفید بودن سیستم با استناد به نمرات امتحان
همکاری فکری	برای ارزیابی همکاری دانشکده با توجه به میزان استفاده از کتابخانه
انتشارات	برای پیدا کردن تأثیر انتشارات در تقاضا برای رشته ها
بازدید از وب سایت	برای تحلیل سوالات دریافت شده در وب سایت دانشگاه و کمک به ایجاد رشته های جدید دانشگاهی

کاربرد داده کاوی در دانشگاه ملی سنگاپور قابل ملاحظه است. در این دانشگاه از ابزارهای داده کاوی برای شناسایی و دسته بندی دانشجویانی که به کلاسهای پیش نیاز برای واحد درسی ارائه شده نیاز داشتند استفاده شد. (Kurian and John, 2005)

علاوه بر آن، مسائلی مانند اختصاص بهتر منابع و نیروی انسانی، مدیریت روابط دانشجو و به تصویر کشیدن رفتار گروههای مختلف میتواند بوسیله ابزارهای داده کاوی انجام شود.

داده کاوی و مدیریت بهینه وب سایت ها

هر سایت اینترنتی بر اساس حجم فعالیت خود برای نگهداری به افراد مختلفی که آشنا به امور فنی و اجرایی باشند نیاز دارد. مدیر سایت به عنوان شخصی که تنظیم کننده و هماهنگ کننده تمام این افراد است باید برای هر کدام از بخش‌های سایت از قبیل گرافیک، محتوا، امور فنی، بازاریابی و... برنامه‌های مختلفی را تهیه و برای اجرا در اختیار همکاران خود قرار دهد. این برنامه ها می توانند شامل برنامه های روزانه، هفتگی و ماهانه باشند. تمامی این برنامه ها در راستای یک هدف کلی و نهایی به انجام می رسند و آن هم بالا رفتن کارایی اقتصادی سایت است.

سایت ها زمانی می توانند خود را در سطح اقتصادی اطمینان بخشی قرار دهند که از بازدیدکنندگان و کاربران و قابل توجهی برخوردار باشند. برای این کار مدیر سایت سعی می کند مطالعه و تحقیق گسترده ای بر روی عوامل و ابزارهای افزایش دهنده تعداد کاربران سایت انجام دهد و از این طریق در واقع به مطالعه شرایط و موقعیت خود در بازار مجازی اینترنت می پردازد. به عنوان مثال وی در مورد

رنگ های به کار رفته در سایت، لوگو و سایر قطعات گرافیکی سایت، متن های به کار رفته و بسیاری دیگر از مسائل مرتبط با سایت به بررسی و مطالعه می پردازد.

یکی از روش ها و راهکارهایی که کمک بسیار زیادی برای بهتر شدن فرآیند مدیریت وب سایت ها می کند استفاده از گزارش ها و تحلیل های آماری است. مدیران سایت ها و مدیران بازایابی شرکت ها با استفاده از گزارش های به دست آمده از فعالیت سایت اینترنتی می‌توانند شناخت خوبی از موقعیت و تاثیر فعالیت های خود پیدا کنند و از این طریق نقاط ضعف و قوت سایت را به راحتی شناسایی و برای حل و تقویت آنها تغییرات لازم را در سایت اعمال نمایند و به برنامه های آینده و حتی استراتژی های سایت جهت ببخشند.

داده‌کاوی و مدیریت دانش

اگر چه دانش به طور انحصاری محصول فناوری اطلاعات نیست، ولی فناوری اطلاعات به طور لاینفکی در ایجاد دانش و فرآیند مدیریت دانش از سال های اول مشارکت داشته است. امروزه مدیریت دانش از مسئولیت های فناوری اطلاعات به شمار می‌رود. زیرا در جمع‌آوری، تبدیل دانش و انتقال داده‌ها، اطلاعات و دانش نقش کلیدی دارد.

از منظر مدیریت دانش، هدف داده‌کاوی، کشف دانش سازمانی پنهان در اطلاعات خام است. اینگونه نیست که هر بینش حاصل از داده‌کاوی دانش می‌سازد، بلکه در عوض بسیاری از نتایج به دست آمده، اطلاعات مدیریت، یا هوش سازمانی است. مثلاً در سازمان‌های تجاری، دانش با ارزش مورد مشتری، محصول و بازار را می‌توان از طریق داده‌کاوی به دست آورد. داده‌کاوی ابزار مفیدی برای مدیران دانش است که کشف را با تحلیل تلفیق می‌کنند. تلفیقی که اغلب منجر به ایجاد دانش می‌شود.

کاربرد داده‌کاوی در آموزش عالی

از هنگامی که رایانه در تحلیل و ذخیره سازی داده ها بکار رفت (۱۹۵۰) پس از حدود ۲۰ سال، حجم داده ها در پایگاه داده ها دو برابر شد. ولی پس از گذشت دو دهه و همزمان با پیشرفت فن آوری اطلاعات (IT) هر دو سال یکبار حجم داده ها، دو برابر شد. همچنین تعداد پایگاه داده ها با سرعت بیشتری رشد نمود. این در حالی است که تعداد متخصصین تحلیل داده ها و آمارشناسان با این سرعت رشد نکرد. حتی اگر چنین امری اتفاق می افتاد، بسیاری از پایگاه داده ها چنان گسترش یافته اند که شامل چندصد میلیون یا چندصد میلیارد رکورد ثبت شده هستند و امکان تحلیل و استخراج اطلاعات با روش های معمول آماری از دل انبوه داده ها مستلزم چند روز کار با رایانه- های موجود است. حال با وجود سیستم های یکپارچه اطلاعاتی، سیستم های یکپارچه بانکی و تجارت الکترونیک، لحظه به لحظه به حجم داده ها در پایگاه داده های مربوط اضافه شده و باعث به وجود آمدن انبارهای (توده های) عظیمی از داده ها شده است به طوری که ضرورت کشف و استخراج سریع و دقیق دانش از این پایگاه داده ها را بیش از پیش نمایان کرده است (چنان که در عصر حاضر گفته می شود « اطلاعات طلاست»).

هم اکنون در هر کشور، سازمان ها، شرکت ها و . . . برای امور بازرگانی، پرسنلی، آموزشی، آماری و . . . پایگاه داده ها ایجاد یا خریداری شده است، به طوری که این پایگاه داده ها برای مدیران، برنامه ریزان، پژوهشگران و . . . جهت تصمیم گیری های راهبردی، تهیه گزارش های مختلف، توصیف وضعیت جاری خود و . . . می تواند مفید باشد. داده کاوی[۱۸] یا استخراج و کشف سریع و دقیق اطلاعات با ارزش و پنهان از این پایگاه داده ها از جمله اموری است که هر کشور، سازمان و شرکتی به منظور توسعه علمی، فنی و اقتصادی خود به آن نیاز دارد.

در کشور ما نیز سازمان ها، شرکت ها و مؤسسات دولتی و خصوصی به طور فزاینده ولی آهسته در حال ایجاد یا خرید نرم افزارهای پایگاه داده ها و مکانیزه کردن سیستم های اطلاعات خود هستند، همچنین با توجه به فصول دهم و یازدهم قانون برنامه سوم توسعه در خصوص داد و ستدهای

[18] - Data mining

الکترونیکی و همچنین تأکید بر برخورداری کشور از فن آوری های جدید اطلاعات برای دستیابی آسان به اطلاعات داخلی و خارجی، دولت مکلف شده است امکانات لازم برای دستیابی آسان به اطلاعات، زمینه سازی برای اتصال کشور به شبکه های جهانی و ایجاد زیر ساخت های ارتباطی و شاهراه های اطلاعاتی فراهم کند. واضح است این امر باعث ایجاد پایگاه های عظیم داده ها شده و ضرورت استفاده از داده کاوی را بیش از پیش نمایان می سازد.

با توجه به اینکه آموزش عالی همواره با داده‌ها و اطلاعات بسیار زیادی در مورد دانشگاه‌ها، دانشجویان، اعضای هیئت علمی، پرسنل، منابع مادی و... روبروست و در اکثر مواقع این داده‌ها می‌تواند حامل اطلاعات و الگوهای باارزشی باشند، لذا به نظر می‌رسد یکی از مهمترین کاربردهای داده‌کاوی در آموزش عالی است. امروز بانک‌های اطلاعاتی وسیعی از ویژگی‌های دانشجویان موجود است که اطلاعات مربوط به ویژگی‌های خانوادگی، تحصیلی و ... را شامل می‌شود. پیدا کردن الگوها و دانش نهفته در این اطلاعات می‌تواند به تصمیم‌گیرندگان عرصه آموزش عالی کمک شایانی بکند. استفاده از تکنیک‌های پیشرفته داده‌کاوی مانند خوشه‌بندی، طبقه‌بندی، و ... می‌تواند در طبقه‌بندی دانشگاه‌ها، یافتن الگوهای خاص و با ارزش در مورد دانشجویان موفق، یافتن یک برنامه یا روش موفق تدریس، یافتن نقاط بحرانی در مدیریت مالی دانشگاه‌ها و موارد دیگر کاربرد داشته باشد.

فصل سوم – بررسی موردی ۱: وب کاوی

معماری وب کاوی

این عامل هوشمند ، در خصوص هر یک از چالش های رسم الخط زبان فارسی رایانه ای ، رفتار متفاوتی از خود نشان می دهد. این رفتارها به قرار زیر است :

الف) تنوع نحوه استفاده از "می" ، "ها" ، پیشوند ها و پسوند ها :

همانطور که قبلا توضیح داده شد ، موارد فوق بطور چسبیده یا جدا از کلمه بکار برده می شود. لذا جهت رفع چنین مشکلی ، می توان در واسط هوشمند ، با حذف کلیه فواصل خالی(Blanks) موجود در عبارت مورد کاوش ، اقدام به جستجو بر اساس دنباله ای از حروف همان عبارت ، بدون هیچگونه فاصله خالی نمود.

ب) بکاربردن "حمزه" بصورت های مختلف :

جهت حل مشکل فوق ، در عمل هوشمند مورد بحث ، فرآیندی ایجاد می گردد ، که طی آن ، اگر عبارت مورد کاوش حاوی صور مختلف "حمزه" باشد ، عملا کاوش ، به چندین جستجو برای کلمات مشابه ، با حالت های مختلف "حمزه" تبدیل می شود. بعبارت دیگر کاوش کلمه "مسئله " به کاوش برای کلمات "مسئله" و "مسأله" منجر می شود. می توان با جایگزینی "ی" بجای "ء" نیز دامنه کاوش را وسیع تر نمود ، مثل "رئیس" و "رییس".

ج) استفاده یا عدم استفاده از "ء" در ترکیب های اضافی یا وصفی :

جهت رفع این مشکل ، در صورت استفاده کاربر از "ء" در عبارت مورد کاوش خود ، واسط هوشمند اقدام به جستجو برای عبارتی فاقد "ء" می نماید. در این صورت نتایج جستجو ، صفحاتی را که در

محتوای متن آنها از "ء" استفاده نشده است نیز شامل می گردد.

د) استفاده از "ا" و "آ":

در این مورد ، واسط ، بمحض برخورد به کلمه مورد کاوش که در آن "ا" بصورت چسبان یا غیرچسبان بکار رفته باشد یا شامل "آ" باشد ، جستجو را به کاوش برای کلمات جدیدی که با جایگزینی "ا" با "آ" ویا "آ" با "ا" ، ساخته شده اند ، بسط می دهد. در نتیجه کاوش برای کلمه "فرایند" ، صفحات حاوی کلمه "فرآیند" ، از دست نمی رود.

ه) استفاده از اصطلاحنامه(Thesaurus) برای حل مشکل تنوع املایی کلمات :

این معضل شامل تنوع استفاده از "ی" در کلمات عربی مختوم به "ا" ، تنوع املایی بعضی کلمات که همه درست هستند ، استفاده از کلمات اروپایی بصورت ترجمه فارسی و استفاده یا عدم استفاده از جمع مکسر برای بعضی کلمات می باشد که حل مشکل کلیه موارد ، در ایجاد یک پایگاه داده در سمت خدمت گذار ، مستتر است. این پایگاه داده شامل نمایه ای از این کلمات و کلمات مترادف می باشد. برای مثال کلمه "موسی" ، به کلمه "موسا" و کلمه "کامپیوتر" به کلمه "رایانه" متناظر شده است. عامل هوشمند با مراجعه به این پایگاه داده ، برای عبارت مورد کاوش کاربر ، عبارات مشابهی استخراج کرده ، کاوش را به جستجو برای این عبارات ، علاوه بر عبارت اصلی ، بسط می دهد.

ایجاد چنین پایگاه داده ای ، با مشاوره انجمن ها ، بزرگان و فرهنگستان ادب فارسی انجام می پذیرد و بروزآوری آن نیز بصورت دوره ای و با دخالت صاحب نظران مذکور صورت می گیرد.نمونه ای از محتویات این پایگاه داده در جدول زیر آمده است :

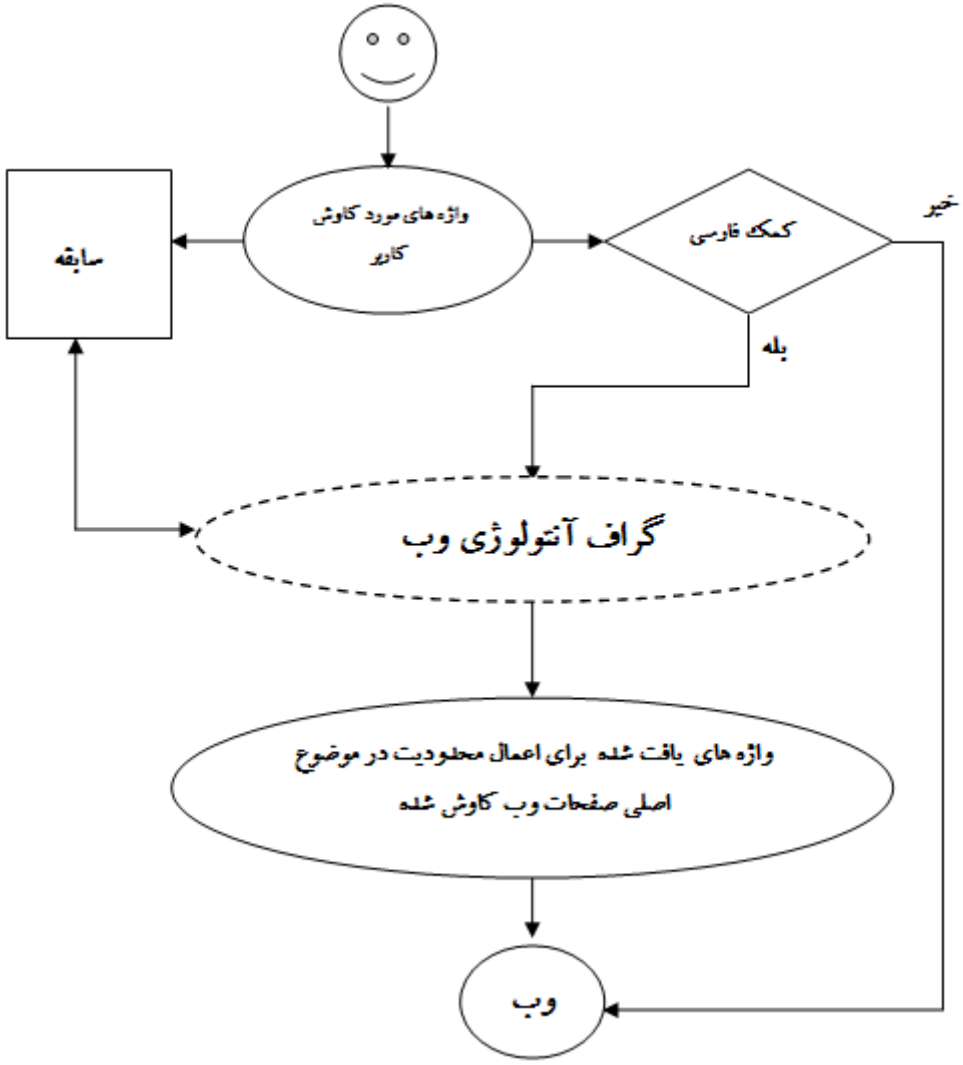
شناسه	واژه اصلی	واژه مترادف
۱	موسی	موسا
۲	امپراتور	امپراطور
۳	Ontology	آنتولوژی
۳	آنتولوژی	انتولوژی
۳	آنتولوژی	انتالوژی
۳	آنتولوژی	هستی شناسی
۴	کامپیوتر	رایانه
۴	Computer	کامپیوتر
۵	Source	منبع
۵	Source	سورس

جدول (۴) نمونه ای از محتویات پایگاه داده مترادف ها.

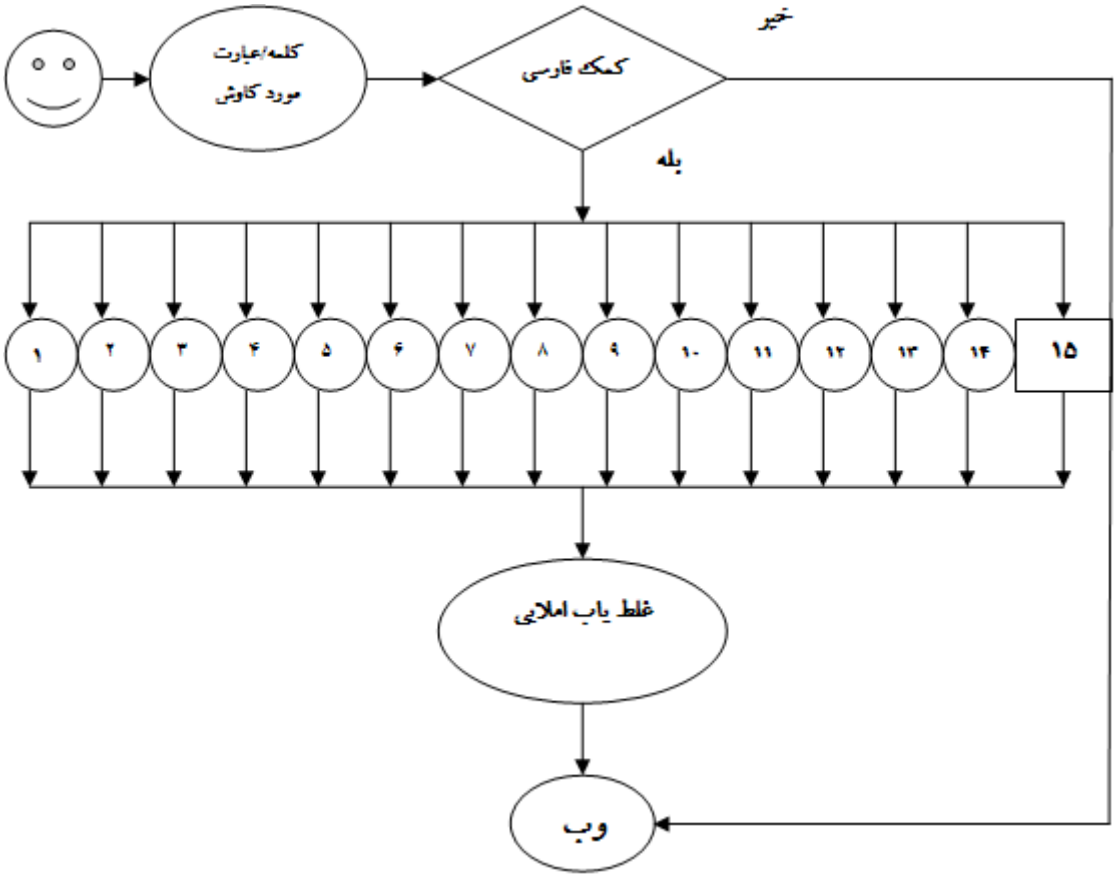
و) تبدیل کلمات اروپایی به رسم الخط فارسی با همان تلفظ اصلی(language Retrieval Cross) :

کاربری که بدنبال اطلاعاتی در خصوص برنامه های "Open Source" در اینترنت می باشد ، شاید برای همیشه از دسترسی به صفحاتی که در آنها کلمه "سورس باز" بکار رفته است ، محروم بماند یا حداقل محکوم به اتلاف زمان زیادی تا رسیدن به چنین کلمه ای و به تبع ، نتایج مطلوب باشد. لذا در صورتی که جستجو برای لغت "سورس" ، بنحوی همزمان با کاوش برای کلمه "Source" ، حتی بدون اطلاع کاربر ، انجام پذیرد ، می توان گفت هم در سرعت و هم در جامعیت اطلاعات بدست آمده ، ارتقایی صورت گرفته است.

وظیفه واسط ما در این خصوص اینست که با مراجعه به پایگاه داده ، کاوش را به کلمه ساخته شده بر اساس تلفظ انگلیسی متناظر نیز گسترش دهد. برای انجام فرآیند حل این مشکل بصورت اتوماتیک و ضمنا استفاده از پایگاه داده معتبرتر و روزآمدتر بعنوان معیار عملکرد این واسط ، می توان روشی پیشنهاد نمود که کلمه متناظر تلفظ انگلیسی لغات که با رسم الخط فارسی تهیه می گردد ، با مراجعه به پایگاه های داده بین المللی حاوی معادل های سمبولیک تلفظ کلمات انگلیسی (که در کتاب های دیکشنری انگلیسی به انگلیسی آمده است) ، کلمه مذبور را تهیه نمود و سپس کاوش را برای آن انجام داد.



شکل (۳) ساختار واسط کمک فارسی برای بهبود مانعیت



شکل (۴) ساختار واسط کمک فارسی برای بهبود جامعیت

نام جزء	پردازش مربوط	نام جزء	پردازش مربوط
C1	حذف "ء" از عبارت	C7	تبدیل "آ" به "ا"
C2	تبدیل "ؤ" به "ئ" و بالعکس	C8	تبدیل "ا" به "آ" در ابتدای کلمات
C3	تبدیل "ئ" به "أ" و بالعکس	C9	اضافه کردن "ء" به "ه" در ترکیبات
C4	تبدیل "ؤ" به "أ" و بالعکس	C10	حذف اعراب ها
C5	تبدیل "یی" به "ئی" و بالعکس	C11	تبدیل "ۀ" به "ه ء" یا به "ه" و بالعکس
C6	تبدیل "ی "به کاراکتر با یونیکد مشابه	C12	مراجعه به پایگاه داده واژه های مترادف

جدول (۵) اجزاء وپردازش های مربوط به رفع اشکالات رسم الخط

مشکلات ومحدودیت های وب کاوی در سایت های فارسی زبان

در دهه های اخیر ، بیشترین اختلاف نظر در باب شیوه املای کلمات فارسی بر سر موضوع جدانویسی یا پیوسته نویسی کلمات مرکب بوده است.فرهنگستان زبان و ادب فارسی ، در این باب راه میانه را برگزیده و کوشیده است تا فقط مواردی را که جدانوشتن و یا پیوسته نوشتن آنها الزامی است ، تحت قاعده و ضابطه درآورد و شیوه نگارش بقیه کلمات مرکب را به ذوق و سلیقه نویسندگان واگذار کند.]فرهنگستان،۱۳۸۲]

بعضی چالش های زبان فارسی در رایانه و بخصوص در اینترنت که باعث تفاوت در نتیجه جستجو در وب یا وب کاوی می شود از قرار زیر است:

الف) تنوع نحوه استفاده از "می" چسبان و غیر چسبان ، مثل کلمات "می تواند" و'' میتواند.''

ب) تنوع نحوه بکاربردن چسبان و غیر چسبان "ها" ، مثل "آن ها" و "آنها.''

ج) بکار بردن بعضی پیشوند ها و پسوند ها ، مثل "همین که" و "همینکه" ویا "هیچ یک" و'' هیچیک" و یا "راه گشا" و "راهگشا.''

د) بکاربردن "حمزه" بصورت های مختلف ، مثل "مسؤول" و "مسئول" یا "مسأله" و "مسئله.''

ه) استفاده یا عدم استفاده از "ء" ، برای کلمات مختوم به های بیان حرکت ، در حالت مضاف ، مثل "خانهٔ مسکونی" و "خانه مسکونی.''

و) تنوع استفاده از "ی" در کلمات عربی مختوم به "ا" ، مثل "موسی" و "موسا.''

ز) تنوع املایی بعضی کلمات که همه درست هستند ، مثل "اتاق" و "اطاق.''

ح) استفاده از کلمات اروپایی بصورت زبان اصلی یا ترجمه فارسی بخصوص در متون علمی ، مثل "Update''و "بروزآوری.''

ط) استفاده یا عدم استفاده از جمع مکسر برای بعضی کلمات.

ی) تبدیل کلمات اروپایی به رسم الخط فارسی با همان تلفظ اصلی ، مثل "Source" و" سورس."

ک) استفاده از "ا" و "آ" بجای هم ، مثل "فرایند" و "فرآیند."

ل) استفاده یا عدم استفاده از اعراب برای کلمات.

بعبارت دیگر ، یک کاربر ممکن است در جستجوی خود در وب ، کلمه کلیدی خاصی را بکار برد ، لیکن در صفحات وب چنین کلمه ای بکار نرفته باشد و با توجه به مواردی که در مورد تنوع کاربری کلمات ، بحث شد ،کلمه مشابهی ثبت شده باشد. بنابراین بسیاری از صفحات وب مطلوب کاربر ، در مجموعه بازیابی شده ، وجود نداشته باشد

محتوا کاوی وب

محتواکاوی وب (Web Content Mining)، یکی از سه شاخه وب کاوی است که در واقع ، کشف اطلاعات مفید از مستندات و داده های ساختیافته و نیمه ساختیافته و غیر ساختیافته وب می باشد. یک شاخه دیگر این مقوله ، ساختارکاوی وب (Web Structure Mining)است که به کشف مدل پشت زمینه حاکم بر ساختار فرا پیوند های وب می پردازد و هدف آن ، ایجاد اطلاعاتی همچون تشابه یا ارتباط بین سایت های مختلف وب است. شاخه دیگر آن کاربرد کاوی وب می باشد که سعی می کند از تعاملات کاربربا وب ، اطلاعاتی کسب کند و از آن ها بصورت سابقه ای در مراجعات بعدی کاربر سود ببرد.

در زمینه محتواکاوی وب نرم افزارهای خزنده (Crawler)، به گشت و گذار در اقیانوس وب پرداخته ، اقدام به نمایه سازی واژگان در پایگاه داده خود می نمایند که مورد استفاده موتورهای کاوش ، در

زمان جستجوهای کاربر قرار می گیرد. نمونه بارز این روش ، موتور کاوشگر Google است .
[Chakrabarti,1999].

در همین راستا ابزارهایی همچون FASTUS:Finite-State Automaton Text Understanding System، در خلال این ماموریت به تجزیه و تحلیل متون ، با هدف کشف گروه های مختلف واژگان مانند اسامی ، افعال ، ترکیبات وصفی و اضافی ،... می پردازند که به کشف دانش از محتویات وب کمک می کند. این روش هم اکنون برای زبان های انگلیسی و ژاپنی پیاده سازی شده است وبصورت بالقوه برای دیگر زبان ها قابل استفاده است[Feiyu,2001] .

از طرف دیگر استفاده از آنتولوژی (Ontology)در وب در بهینه سازی کاوش در وب پیشنهاد می گردد. آنتولوژی ، یک فرهنگ واژگان مشترک بر اساس موضوع سایت برای استاندارد سازی ارائه مفاهیم آن جهت قابل تفسیر شدن توسط ماشین ، تعریف می کند. آنتولوژی ، یک جزء کلیدی وب مفهومی (Semantic Web)است[Heflin,2000] .

شخصی کردن وب (Personalization)، از دیگر روش هاست که در امر کاوش وب مثمر ثمر است. نمونه این روش در My Yahoo قابل مشاهده است.

یکی دیگر از راه های کاوش در مقدار زیاد و غیر ساختیافته اطلاعات وب ، استفاده از پایگاه داده چند لایه ای (MLDB) است. هر لایه از این پایگاه داده ، تعمیم بیشتری از لایه قبلی است. همه لایه ها بجز پایین ترین لایه) که خود وب است) ، قابل کاوش توسط یک زبان پرس وجو مثل SQL است .
[Osmar,2002]

در پیاده سازی روش های ساختارکاوی وب ، از تئوری گراف وب بهره مند خواهیم شد که به ایجاد دید ارزشمند در الگوریتم های جستجو ، کشف ارتباطات ،... موثر است.

در خصوص روش های کاربرد کاوی وب ، ناوبری کاربر در وب توسط مدل های ریاضی

مارکو (Markov)، براساس میزان تجربه کاربر و دارا بودن یا عدم داشتن راهنمای سایت ، تجزیه و تحلیل می گردد .

فصل چهارم – بررسی موردی ۲ : داده کاوی در شهر الکترونیک

شهرالکترونیکی[۱۹] یکی از دستاوردهای موج سوم است. با گسترش صنعت ارتباطات و سپس فراگیر-شدن اینترنت، فضای جدیدی برای شهرها به‌وجودآمده است که از آن به شهر الکترونیکی یا شهر مجازی، تعبیرمی‌شود. تحولات حوزه فناوری بویژه گسترش شبکه اینترنت در دهه اخیر و انقلاب اطلاعاتی، موجب تحولات عمیقی شده و جوامع انسانی را به‌سوی تبدیل‌شدن هر چه بیشتر و سریعتر، به جوامع الکترونیک پیش‌می‌برد. جوامعی که درآنها، مرزهای میان واقعیت و خیال تا اندازه زیادی مخدوش شده و یا به عبارت دیگر، وارد رابطه‌ای کاملا تازه شده است. اینترنت، به‌نوعی آغاز ساختمان شهری تازه است. شهربالقوه، شهری متعلق به زمان واقعی، شهری با یک شبکه ارتباطی درونی براساس سرعت مطلق امواج الکترومغناطیسی نوعی فرامرکز شهر جهان است که تمام شهرهای واقعی اقمار و حاشیه این شهر بالقوه‌اند و این شهر بالقوه درهیچ‌جا نیست. تنها در آنجاست که آنتن‌های بشقابی و فیبرهای نوری در آنجا هستند. شهر مجازی، شهری نیست که در مکان واقعی قرارگرفته‌باشد. این شهر به مکان–زمان جغرافیایی وابسته نیست بلکه با زمان واقعی پخش و دریافت فوری یک پیام، رابطه دارد . در چنین شرایطی، هویت شهری و حتی شهروندی نیز معانی متفاوتی می‌یابند.

مطابق تعریف کارشناسان، شهر الکترونیکی شهری است که در آن از ابزار ICT نظیر برنامه‌های کاربردی و کامپیوتر برای افزایش کارایی و اثر بخشی خدمات به مردم، بنگاه‌های اقتصادی وکارمندان دیگر بخش‌های دولت استفاده‌می‌شود [۱]. به‌طورکلی می‌توان مزایای بالقوه شهر الکترونیک را برای شهروندان، کسب‌وکارها و موسسات اداری عمومی، به صورت زیر خلاصه‌کرد:

---

[۱۹] Electronic City(E-city)

– افزایش دسترسی: شهروندان و کسب‌وکارها می‌توانند از طریق کانال‌های مبتنی بر تکنولوژی‌های اطلاعاتی و ارتباطاتی از قبیل کیوسک‌های اطلاعاتی، اینترنت و دستگاه‌های موبایل، با دولت تعامل داشته‌باشند.

– انعطاف‌پذیری: توانایی تعامل در زمان‌های مناسب‌تر و تولید انواع سرویس‌ها از ویژگی‌های شهر الکترونیک محسوب می‌شود.

– کارایی: یک دولت کارامدتر در شهر الکترونیک، منجر به خدمات بهتر و استفاده بهتری از منابع موجود می‌شود.

– پوشش بیشتر: توانایی دسترسی به درصد بیشتری از افراد جامعه مثل مناطق روستایی و محروم، افراد معلول و بیکار و غیره.

شهر الکترونیک، از شایسته‌ترین کاربردهای داده‌کاوی[۲۰] به حساب‌می‌آید. در متون آکادمیک تعاریف گوناگونی برای داده‌کاوی ارائه شده‌است. در واقع داده‌کاوی روشی برای پشتیبانی تصمیم گیری مبتنی بر کامپیوتر است و با گرفتن الگوریتم‌های زیادی از آمار، هوش مصنوعی و سایر زمینه‌ها کاری جدید را انجام می‌دهد. نقطه تحول الگوریتم‌ها، داده کاوی نیست بلکه، ایده استخراج دانش به صورت خودکار از پایگاه داده‌های بزرگ است .

با گسترش سیستم‌های پایگاهی و حجم بالای داده‌های ذخیره شده در سیستم‌های شهر الکترونیک، نیاز به ابزاری است تا بتوان داده‌های ذخیره شده را پردازش‌کرد و اطلاعات حاصل از این پردازش را در اختیار دولت و کاربران قرارداد. یکی از این ابزارها، داده‌کاوی است که همزمان با ایجاد و استفاده از پایگاه داده‌ها در اوایل دهه ۸۰ برای جستجوی دانش در داده‌ها شکل گرفت. این داده‌ها برای ثبت رفتار کاربران (شهروندان و سازمان‌ها)، اهداف و انگیزه‌های کاربران بکارمی‌روند (شکل شماره ۱).

---

[۲۰] Data Mining(DM)

بهره‌برداری از چنین داده‌هایی دولت را در شناسایی بیشتر شهروندان و سازمان‌ها، نیازمندی‌ها و رفتار-های آنان، کشف تخلفات و اتخاذ سیاست‌های مناسب، یاری‌می‌دهد.

در سال‌های اخیر تحقیقات روبه‌رشدی در حوزه داده‌کاوی توسط محققین انجام شده‌است. از نقطه‌نظر شهر الکترونیک، داده‌کاوی این پتانسیل را دارد که موجب کاهش هزینه‌ها و کسب مزایای رقابتی برای همه سهام‌داران یعنی شهروندان و کسب‌وکاها شود.

در حالی که ازدیاد کاربردهای داده‌کاوی می‌تواند ابزارهای مدیریتی داده بسیاری عرضه کند، سازمان‌هایی که این ابتکارات را استعمال می‌کنند به محدودیت‌های سوتدبیر پیاده‌سازی آن و اشتباهات و غفلت‌های آن آسیب‌پذیر هستند. به عبارت دیگر درحالی‌که ابتکارات داده‌کاوی رو به تکامل است، مسائل و چالش‌هایی در رابطه با پیاده‌سازی و بررسی آن وجوددارد که جوامع سیاسی و دولتی برای جلوگیری از خطرات زیان‌بار بایستی از آنها آگاه‌باشند.

شکل (۱): جایگاه داده‌کاوی در شهر الکترونیک

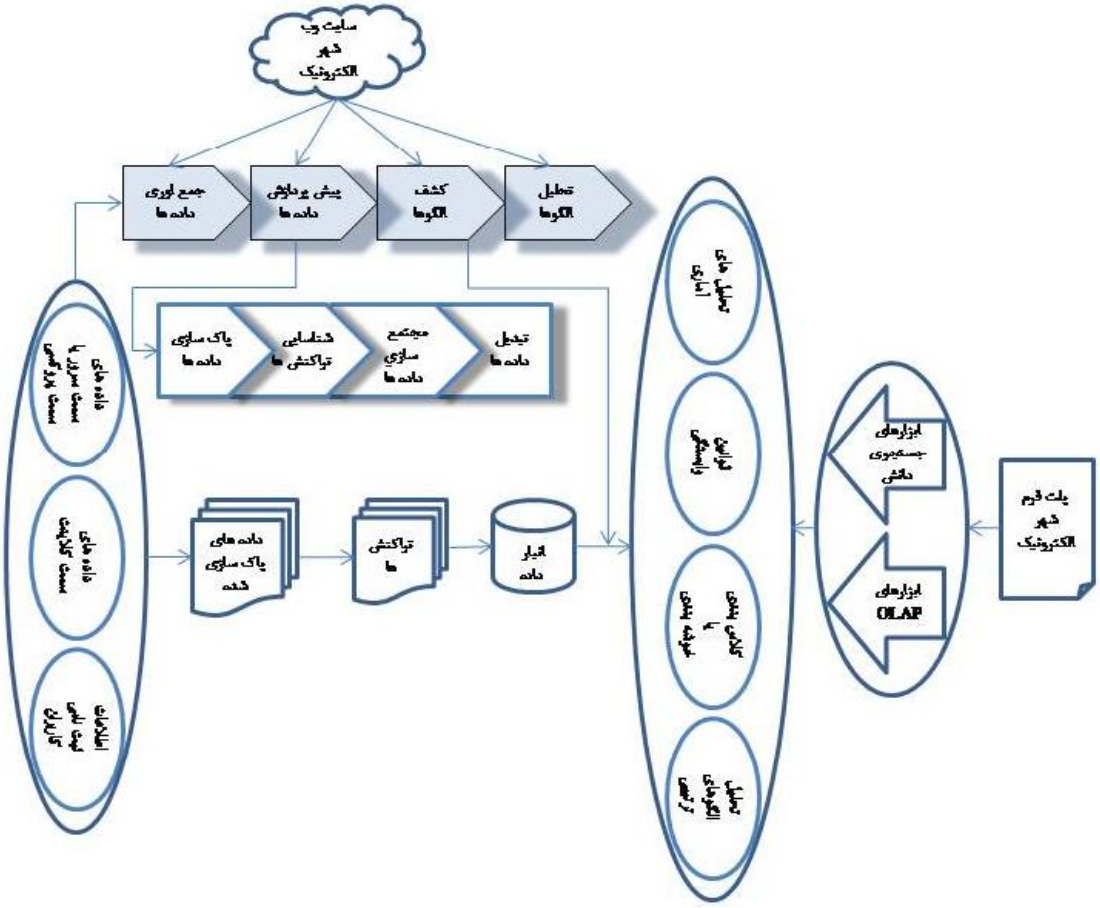
زمینه داده‌کاوی در شهر الکترونیک

درحالی‌که پردازنده‌های مرکزی بزرگ و گران‌قیمت در دهه ۶۰ و ۷۰ میلادی به منظور پردازش کنترل موجودی‌ها و مدیریت داده‌ها برنامه‌نویسی شدند، اکثر هزینه‌ها به‌سادگی بر روی دستگیری داده به فرمت قابل خواندن برای ماشین، و تحویل آن به سیستم‌ها متمرکز شد، به‌طوری که متجاوز از ۸۰

درصد هزینه پروژه‌های نرم‌افزاری در دوره دهه ۱۹۶۰ تا اوایل دهه ۹۰ صرف‌شده‌است. تاکید روی مسایل تمامیت داده‌ها و کنترل صحت داده‌ها، در اواخر دهه ۹۰ به نقطه اوج خود رسید و در حالی که سال ۲۰۰۰ سپری می‌شد این مسئله منجر به سرمایه‌گذاری بسیاری از سازمان‌ها بر روی ایجاد پایگاه‌های داده به‌منظور جلوگیری از خطاهای فاجعه‌انگیز شد. یک واکنش ثانویه حرکت به‌سمت اجتناب از مشکلات Y2K، این بود که شرکت‌ها اقدام به پیاده‌سازی سیستم‌های مدیریت پایگاه داده[۲۱] گران-قیمتی کردند که توانایی ضبط و نگهداری، پردازش و اشتراک بهتر داده‌ها را داشته‌باشد.

با وجود این موهبت تازه و ارزان، داده‌های صحیح، بروز، قابل دسترس و قابل پردازش از سیستم‌های اطلاعاتی، پس از سال ۲۰۰۰ هزینه‌های گسترش کاربردهای داده‌کاوی کاهش یافت و متدهای جدیدی از دستگیری صریحانه دانش یا هوش از انبارهای داده‌ای بزرگ و توسعه‌پذیر شکل گرفت. نتیجه اینکه متدهای داده‌کاوی نه فقط برای انبار داده شخصی یک سازمان، بلکه بر روی داده‌های کپی‌شده، خریداری‌شده و یا حتی از منابع خارجی یا دیگر ادارات دولتی به‌سرقت‌رفته نیز قابل استفاده بود .

[۲۱] Database Management System(DBMS)

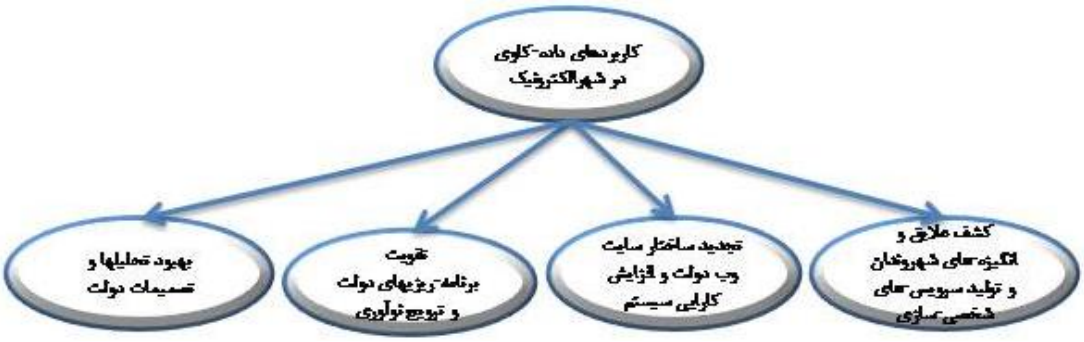


شکل (۲): چهارچوب داده‌کاوی در شهر الکترونیک

کاربردهای داده‌کاوی در شهر الکترونیک

در این بخش، کاربردهای مهم داده‌کاوی در شهر الکترونیک ارائه شده‌است (شکل شماره ۳). این کاربردها، پروسیجرهایی هستند که داده‌های به‌دست‌آمده از تعامل شهروندان یا سازمان‌ها با سیستم‌های الکترونیکی شهری را به منظور پشتیبانی تصمیم‌های شهر در امور مختلف، به دانش با ارزش تبدیل می‌کند. این پشتیبانی‌های تصمیم، شامل یافتن علایق، انگیزه‌ها و خواسته‌های شهروندان و بهبود رضایت شهروندان و سازمان‌ها، سازمان‌دهی مجدد سایت وب شهر و افزایش عملکرد و کارایی سیستم، تقویت

برنامه‌ریزی‌های خدمات شهری و ترویج نوآوری و بهبود تحلیل‌ها و تصمیم‌گیری‌ها می‌شود. در ادامه به تفصیل هر یک از موارد مذکور می‌پردازیم.



شکل (۳): کاربردهای داده‌کاوی در شهر الکترونیک

کشف علایق و انگیزه‌های شهروندان و تولید سرویس‌های شخصی‌سازی

فعالیت‌های کاربران می‌تواند نمایانگر رفتارهای آنان و علایق آنها باشد. مثلا کجاها را کلیلک می-کنند، چه‌مدت روی یک صفحه خاص توقف می‌کنند، چه کلماتی را جستجو می‌کنند، از چه سایت وبی مراجعه می‌کنند، روابط متقابل انها با سایت وب چیست و از این قبیل. سرویس‌های شخصی-سازی[۲۲]، یک کاربرد مبتنی بر وب و یک هدف نهایی است که بر مبنای داده‌های ثبت‌نام کاربران و داده‌های به‌دست‌آمده از تعامل آنها با سایت وب، به یافتن خوشه‌هایی از کاربران با الگوهای دستیابی مشابه و سلایق یکسان می‌پردازد. آن‌گاه هر کاربر را با توجه به الگوی پیمایشی فعلی، به یک خوشه واحد منسوب می‌کند و سپس پیشنهاداتی برای بقیه صفحات درون سایت وب، به او عرضه می‌کنند. این پیشنهادها بر اساس اینکه چه صفحاتی توسط کاربران همان خوشه رویت شده‌است، به صورت پویا تولیدمی‌شوند . کل این فرایند در شکل ۴ به نمایش درآمده‌است

[۲۲] Personalization Service

.

شکل (۴): فرایند تولید سرویس‌های شخصی

بنابراین در شهر الکترونیک، ثبت کردن داده‌های کاربرانی که در سایت شهر حرکت‌می‌کنند و شناسایی لینک‌هایی که به‌طور بالقوه مورد علاقه آنها است، به‌منظور یافتن رفتار شهروندان، پی‌بردن به نیازهای آنان و تولید سرویس‌های شخصی شهروندان و مدیریت نیازهای آنان، امری ضروری می‌باشد.

تجدید ساختار سایت وب شهر و افزایش کارایی سیستم

محتوی و ساختار سایت وب شهر همیشه ثابت نیست. طراحان وب‌سایت نمی‌توانند فقط به متخصصین در زمینه طراحی وب‌سایت متکی باشند. آنها بایستی قادر باشند برای سهولت دسترسی شهروندان و سازمان‌ها، ساختار و محتوی سایت وب شهر را بر اساس نتایج به‌دست‌آمده از تحلیل‌های داده‌کاوی در شهر الکترونیک، به صورت پویا تغییردهند. به‌عنوان مثال سایت وب را می‌توان با توجه به مسیرهای دستیابی متوالی بازدیدکنندگان، دوباره‌ساختاردهی کرد. این کار موجب صرفه‌جویی در زمان دسترسی کاربران و هزینه‌ها خواهدشد.

به‌طورکل به‌منظور بهینه‌سازی ساختار سایت وب شهر، پیشنهادات زیر مورد استفاده قرار می‌گیرد :

- کاوش فایل‌های لاگ کاربران به منظور استخراج صفحات دسترسی مرتبط با هم. بدین ترتیب لینک‌های جدید بین صفحات اضافه می‌شود تا حرکت شهروندان را در سایت وب تسهیل سازد.

– استفاده از تکنیک‌های تحلیل مسیر برای یافتن پرتکرارترین مسیرهای دستیابی و ایجاد پیام‌های مهم در بین آنها به‌منظور افزایش تمایل به شهروندان و بهبود کیفیت خدمات.

– کاوش فایل‌های لاگ، می‌تواند موقعیت موردانتظار کاربران را برای اطلاعات آشکارسازد. اگر تعداد تکرار موقعیت موردانتظار دسترسی از مکان واقعی دسترسی بیشتر بود، آن‌گاه برای هدایت بهتر شهروندان، یک لینک جدید در بین صفحات افزده‌می‌گردد.

داده‌کاوی، با مشاهده و تحلیل رفتارها و فعالیت‌های شهروندان در سایت وب شهر الکترونیک، تاثیر چشمگیری در در شهر الکترونیک دارد. و با ثبت دانش در حوزه اجتماعی و انسانی مرتبط با فعالیت شهروندان، منجر به بهبود سیستم می‌شود. به عنوان مثال از آنجا که امنیت یکی از موضوعات مهم برای سایت وب شهر الکترونیک محسوب می‌شود با ثبت الگوهای دسترسی کاربران و تحلیل مسیرهای دستیابی آنها توسط تکنیک‌های داده‌کاوی، می‌توان به‌سرعت و سهولت به مزاحمت‌ها و تهاجم‌ها پی برد و با آنها مقابله کرد، مانند تکنیک‌های داده‌کاوی برای تشخیص کلاهبرداری در کارت‌های اعتباری. محققان نشان‌داده‌اند که بسیاری از این تکنیک‌ها از سیستم‌های مهندسی‌شده و نیز افراد خبره، موفق‌تر عمل‌می‌کنند .

تقویت برنامه‌ریزی‌های دولت و ترویج نوآوری

با بکارگیری تکنولوژی داده‌کاوی، دولت می‌تواند منابع انسانی، منابع اطلاعاتی و نیازمندی‌ها را برای تعدیل روابط بین منابع درونی و بیرونی شهر مدیریت‌کند مثلا مدیریت کل فرایند از طرح‌ریزی برنامه-ها گرفته تا پیاده‌سازی آنها با استفاده از داده‌کاوی. ابزارهای OLAP[۲۳] می‌توانند جریان پروژه را متناسب با منابع اجتماعی بهینه کنند، این، به مقدار زیادی هزینه منابع اجتماعی و جابجایی اطلاعات را

[۲۳] Online Analysis and Process

کاهش‌می‌دهد و سبب ترفیع علمی، آگاهی و هوشمندی برنامه‌ریزی‌ها و خدمات شهری می‌شود. به عبارتی دیگر، یک برنامه‌ریزی دولتی موثر و علمی می‌تواند در سایه کنترل و مدیریت بلادرنگ و با ابزارهای تکنولوژی هوشمند و تجسمی تقویت شود.

کاوش داده‌های موجود در سیستم‌های شهر الکترونیک، به صورت معناداری حساسیت دولت را به همه موضوعات از درخواست‌های روزمره شهروندان گرفته تا رخدادهای تشنجی بهبود می‌بخشد. گذشته بر این کاوش چنین داده‌هایی، موجب حصول سریع و بموقع اطلاعات راجع به خدمات شهری و رویه-های اجتماعی توسط دولت می‌شود. بدین ترتیب مدیریت و نقل و انتقال منابع اجتماعی را سیتماتیک-تر و پویاتر ساخته و توانایی‌های نوآوری دولت را بهبود می‌بخشد.

نوآوری فقط محدود به متد اجرایی و حکومتی و جریان فرایند امورات و خدمات شهری نمی‌شود، و شامل وضع استراتژی‌ها و سیاست‌های عمومی می‌شود و این یک الزام اساسی است که دولت را از سازمانی عملیاتی به سازمانی خدماتی تبدیل می‌کند .

بهبود تحلیل‌ها و تصمیمات دولت

۹۰درصد یک تصمیم بر اساس اطلاعات و ۱۰درصد آن بر اساس مهارت‌های ادراکی فرد صورت می-پذیرد. تصمیمات دولت از طریق تحلیل اطلاعات مربوط به شهروندان از قبیل درخواست‌ها، پیشنهادها و تمایلات آنها اتخاذمی‌شود. بنابراین داده‌های حاصل از تعامل کاربران، منبعی از بینش خاص، اطلاعات، دانش و تجربه را تولید می‌کند که با صحت راهکارهای دولت در مواجه با مشکلات دخیل می‌باشد.

مشارکت شهروندان در مسائل عمومی و اجتماعی، به مهارکردن فعالیت‌های سیاسی و خط مشی کشور کمک‌می‌کند. تعامل شهروندان با سیستم‌های الکترونیکی شهری و کاوش داده‌های حاصل از آن می-

تواند یک برنامه، فعالیت، طرح و رهبری را قانونی و مشروع نماید. از طرف دیگر کاوش این اطلاعات منجر به شناسایی اطلاعات مهم و پنهانی می‌شود که تمام سطوح بخش‌های دولتی را پشتیبانی کرده و منجر به کاهش هزینه‌های پرسنلی و هزینه‌های ناشی از انجام فعالیت‌های سیاسی می‌شود .

چالش‌های داده‌کاوی در شهر الکترونیک

در حالی که ازدیاد کاربردهای داده‌کاوی می‌تواند ابزارهای مدیریتی داده بسیاری برای شهر الکترونیک عرضه‌کند، مسائل و چالش‌هایی در رابطه با پیاده‌سازی و بررسی آن وجوددارد که جوامع سیاسی و دولتی برای جلوگیری از تصمیمات نادرست و عواقب آن، بایستی از آنها آگاه‌باشند.

در ادامه، به بررسی بعضی مشکلات چالش‌برانگیز در حوزه‌های مختلف داده‌کاوی در شهر الکترونیکی می‌پردازیم.

کیفیت داده‌ها

کیفیت داده‌ها بزرگ‌ترین چالش داده‌کاوی می‌باشد. منظور از کیفیت داده, تمامیت[۲۴] و درستی داده‌ها می‌باشد و اینها خود موثر از ساختار و سازگاری داده‌های تحلیل‌شده می‌باشد. از آنجاکه تحلیل‌ها و تصمیمات شهر الکترونیک به تکنیک‌های داده‌کاوی وابسته‌است، بنابراین کارایی دولت تا حدودی به اختلاف دقت موجود در داده‌ها حساس است. وجود رکوردهای تکراری, کمبود استانداردهای خاص برای داده‌ها, بروزرسانی داده‌ها, از همه مهم‌تر اشتباهات کاربران همگی از عوامل تعیین‌کننده اثربخشی و قابلیت اجرای تکنیک‌های پیچیده داده‌کاوی هستند .

[۲۴] Completeness

پاک‌سازی داده‌ها, یکی از روش‌های بهبود کیفیت داده است و شامل عملیات نرمال‌سازی, استاندارد-کردن فرمت داده (مثلا تبدیل همه تاریخ‌ها به فرمت MM/DD/YYYY) و نیز حذف مقادیر تکراری و ناخواسته می‌باشد .

قابلیت انتقال داده‌ها و استفاده از اطلاعات

منظور از قابلیت انتقال[۲۵] داده‌ها, توانایی استفاده از سیستم‌های کامپیوتری یا داده در کارکردن با دیگر سیستم‌ها یا داده‌ها, با استفاده از استانداردها و فرایندهای معمول است. این ویژگی یک بخش مهم از تلاش بزرگ‌تری است که همدستی و میانجی‌گری[۲۶] و اشتراک اطلاعات را از طریق شهر الکترونیک و یا ادارات تامین‌کننده امنیت بهبود می‌بخشد .

در کاربردهای داده‌کاوی قابلیت انتقال پایگاه داده‌ها و نرم‌افزار, برای فعال‌کردن جستجو و تحلیل همزمان در چندین پایگاه داده و تضمین سازش‌پذیری فعالیت‌های داده‌کاوی در امورات مختلف شهری بسیار مهم می‌باشد. بنابراین, زمانی‌که دولت و نهادهای دولتی اقدام به ایجاد پایگاه‌داده‌های مختلف و اشتراک اطلاعات می‌کنند, بایستی در فاز طراحی, موضوعات مربوط به قابلیت داده را برای اطمینان از اثربخشی پروژه‌های داده‌کاوی موردتوجه قراردهند.

چالش برآورد مدل‌های داده‌کاوی

همان‌طور که گفته‌شد سیستم‌های پایگاهی شهر الکترونیک و حجم داده‌های ذخیره شده در آنها، روز-به‌روز در حال گسترش است و بنابراین، نیاز به ابزاری است که بتواند داده‌های ذخیره شده را پردازش-کند و اطلاعات حاصل از این پردازش را در اختیار دولت و کاربران قراردهد.

---

[۲۵] Interoperability

[۲۶] Interagency Collaboration

این مجموعه داده‌های بزرگ، به دلایل مختلف می‌توانند چالشی برای داده‌کاوی محسوب شوند. یک دلیل این است که داده‌ها ممکن است آمیخته‌ای از چندین زیربخش مختلف جامعه با ویژگی‌های خاص باشند. به‌طوری‌که هر کدام از بخش‌ها نیازمند مدل داده‌کاوی یا آماری مجزایی هستند. هر کدام از این مدل‌ها فقط با استفاده از داده‌ها تخمین زده می‌شوند. در بعضی کاربردها، زیربخش‌ها ممکن است خود ناشناخته باشند و این چالشی برای داده‌کاوی به‌حساب می‌آیند.

یک روش معمول برای حل این مشکل که تا مدتی پیش به‌عنوان روشی استاندارد استفاده‌می‌شد بخش‌بندی جامعه داده‌ها به صورت دستی و اعمال یک مدل آماری یا تکنیک داده‌کاوی مجزا برای هر بخش بود. برای مثال، یک شیوه رایج برای ایجاد مدل‌های واکنش در بازاریابی، تقسیم مشتریان هدف بالقوه به چندین بخش مختلف و تخمین پارامترهای ایجاد تکنیک‌های مجزا برای هر بخش، است. در این روش، تخمین پارامترها به صورت اتوماتیک در بین هزاران یا میلیون‌ها مدل مجزای داده-کاوی، مشکل بزرگ در مسائلی با مجموعه داده‌های بسیار پیچیده و بزرگ است .

یک مثال از داده‌کاوی چنین داده‌هایی، داده‌کاوی داده‌های حاصل از ۸۳۳ سنسور ترافیک در یک منطقه شهری از شیکاگو به‌منظور شناسایی الگوهای نابهنجار و غیر عادی ترافیک است. علاوه‌بر داده-های سنسورها، داده‌های نیمه‌ساخت‌یافته درباره شرایط آب‌وهوایی و داده‌های متنی مشخص‌کننده وقایع خاصی چون مسابقات ورزشی نیز در تصمیی‌گیری بکارمی‌روند. راهکاری که برای این منظور درنظرگرفته‌شده بخش‌بندی داده به زیربخش‌های مختلف، یکی برای هر ساعت از روز (۲۴ ساعت)، برای هر روز از هفته (۷ روز)، و دیگری برای هر بخش کوچک از بزرگراه (تقریبا ۲۵۰ بزرگراه) است. و منجر به تولید ۲۵۰*۷*۲۴ یا ۴۲,۰۰۰ بخش مختلف شده‌است. برای هر بخش، پارامترهای مدل کشف تغییر برحسب داده‌های هر قسمت تخمین زده‌می‌شود. در این روش، متجاوز از ۴۲,۰۰۰ مدل آماری مختلف به‌صورت اتوماتیک تولید شد و برای تشخیص تغییرات نابهنجار ترافیک استفاده و

بروز شد. ثابت‌شده‌است که به‌علت سایز داده، پیچیدگی و ناهمگونی آن، این راهکار برای ایجاد مدل‌های کمتر، مورد ترجیح است .

امروزه کاربردهای متنوعی از این روش استفاده کرده‌اند. یک مثال، بازاریابی‌های آنلاین می‌باشد که در آن کمپانی‌های بزرگ به‌صورت جداگانه برای هر مشتری به ایجاد یک مدل آماری می‌پردازند. یا تشخیص ناهنجاری در شبکه با ایجاد مدل‌های جداگانه برای آدرس‌های IPv4 و IPv6 .

همان‌طور که گفته‌شد مشکل اصلی در اینجا براورد پارامترهای یک مدل، به‌هدف کاوش حج عظیمی از بردار ویژگی‌های مختلف نیست بلکه براورد پارامترهایی از تعداد زیادی مدل مختلف می‌باشد به‌طوری‌که هر کدام قادر به کارکردن با انبوه بردارهایی با ویژگی مختلف باشند. دومین مشکل به هنگام بخش‌بندی کردن مجموعه داده‌ها رخ‌می‌دهد. به‌طوری‌که اگر دقت کافی انجام نشود تعداد بسیار بخش‌ها منجر به کاهش دقت کل خواهدشد. بنابراین در کلاس‌بندی داده‌ها باید تعداد بخش‌ها و نیز نرخ کلاس‌بندی نادرست[۲۷] کل را به‌حداقل رساند.

دقت نتایج متدهای داده‌کاوی

اکنون بیشتر از ۸۰ درصد از دانش سازمان‌ها به صورت متن، مستندات و دیگر صورتهای رسانه ای نظیر ویدیو و صدا نگهداری می شود.

متن‌کاوی، یک شاخه از داده‌کاوی و یک حوزه تحقیقی مهیج برای کاوش اطلاعات متنی است و تلاش می‌کند تا مشکل بار زیاد اطلاعاتی را به‌وسیله تکنیک‌های داده‌کاوی، پردازش زبان طبیعی، بازیابی اطلاعات و مدیریت دانایی حل‌کند. متن‌کاوی شامل پیش‌پردازش مجموعه اسناد (طبقه‌بندی

[۲۷] Misclassification Rate

متن، استخراج اطلاعات)، ذخیره‌سازی نتایج میانی (تحلیل پراکندگی، خوشه‌بندی، قوانین وابستگی و غیره) و تجسم فکری نتایج می‌شود.

اکثر سیستم‌های تحلیل متن بر مبنای استخراج دقیق موجودیت‌ها و روابط از اسناد هستند. بااین‌حال دقت این سیستم‌ها در بعضی حوزه‌ها فقط به ۷۰ یا ۸۰ درصد می‌رسد و سطحی از نویز تولید می‌کند که مانع از پذیرش کاربرد سیستم‌های متن‌کاوی توسط کاربران می‌شود. بایستی سیستم‌های استخراج روابط موجود باشند که مستقل از دامنه و زبان قادر به تولید دقتی نزدیک به ۹۸ تا ۱۰۰ درصد و فراخوانی به اندازه ۹۵ تا ۱۰۰ درصد تولید کنند.

ضمنا از آنجایی که این‌گونه سیستم‌ها بایستی در هر دامنه‌ای سازگار باشند باید در هر دامنه‌ای، بدون دخالت انسان و مستقل قادر به عمل کردن باشند. امروزه اغلب سیستم‌های تحلیل متنی با هدایت کاربر استفاده‌می‌شوند. شهر الکترونیک برای حل این مشکل بایستی قادر به تولید سیستم‌هایی باشد که در عین استقلال کامل، با تحلیل مجموعه‌های عظیم به یافته‌های درست و جالبی نیل‌شود که در هیچ سند واحدی در مجموعه، ضبط نشده‌باشد و از قبل نیز شناسایی نشده‌باشد. چنین سیستم‌هایی خواهند توانست از اینترنت برای فیلترکردن حقایقی که از قبل ناشناخته‌است، بهره‌ببرند .

انتظار می‌رود که متن‌کاوی یکی از مهم‌ترین تکنولوزی‌ها در آینده باشد بخصوص زمانی‌که سازمان-های بزرگی چون دولت، نیاز به پاسخ سریع، کارا و معنادار به پست‌های الکترونیکی مشتریان دارند. پس توسعه این سیستم‌ها در سیستم‌های شهر الکترونیک، برای هشداردادن در امور مالی، حوزه‌های ضد فعالیت‌های تروریستی، تجارت آنلاین، آموزش الکترونیک و حتی پزشکی مفیدمی‌باشد. و به شهروندان کمک خواهدکرد سریع‌تر به دانش موردنیاز خود درباره خدمات شهری و مقوله‌های مختلف دست‌یابند.

پیچیدگی و هزینه زمانی

تحلیل‌گران دریافته‌اند که پیچیدگی و زمان‌بر بودن دسترسی به حجم زیاد داده‌های موردنیاز و پردازش آنها توسط بعضی ابزارهای داده‌کاوی، استفاده از این ابزارها را در هر نقطه از زمان و مکان غیر ممکن ساخته است.

وزارت امنیت داخلی ایالات متحده آمریکا در آگوست ۲۰۰۶، به ۱۲ تلاش داده‌کاوی دست‌زد که یکی از آنها سیستم TVIS[۲۸] بود. این سیستم به منظور ایجاد و بهبود اشتراک دانش از خطرات تروریستی بالقوه، به روشی واحد داده‌های زنده تولید شده به وسیله خلبانان را ترکیب می‌کرد. نتایج تحلیل‌ها نشان‌داد که اگرچه این سیستم در یک دوره تناوب دو ساعته کارمی‌کند، کاربران قادر به استفاده روزانه از آن نبوده و فقط دو تحلیل‌گر امکان استفاده همزمان از آنرا دارند. این منجر به اتلاف وقت تحلیل‌گران در زمان جستجو در پایگاه داده‌های مضاعف شد .

از آنجا که شهر الکترونیک و شهروندان روزانه با حجم عظیمی از تراکنش‌های زمان واقعی[۲۹] روبرو هستند مشکل پیچیدگی و هزینه زمانی بعضی تکنیک‌های داده‌کاوی، موجب کاهش پذیرش استفاده زمان واقعی از این سیستم‌ها توسط افراد و روی‌آوردن به سیستم‌هایی با عملکرد ضعیف‌تر می‌شود.

محرمانگی[۳۰] داده‌ها

با وجود تکنیک‌های داده‌کاوی و اشتراک اطلاعات، توجه بسیاری از تحلیل‌گران به پیاده‌سازی محرمانگی و امنیت داده‌ها معطوف‌شده‌است. بعضی کارشناسان پیشنهادکرده‌اند که بعضی کاربردهای ضد تروریسمی داده‌کاوی می‌تواند برای یافتن الگوهای تبهکارانه و مقابله با انواع جرم‌ها مفید باشد. تا

---

[۲۸] Threat Vulnerability Integration System

[۲۹] Real Time

[۳۰] Privacy

کنون، با وجود دیدگاه‌های متضاد بحث‌شده، توافق کمی درباره اینکه داده‌کاوی به چه صورت باید اجرا شود وجوددارد. بعضی مخالف سبک‌سنگینی برای ایجاد محرمانگی و تامین امنیت هستند. بعضی مشاهده‌گران نیز پیشنهادکرده‌اند که قوانین و مقررات مربوط به حمایت از محرمانگی کافی هستند و هیچ تهدیدی برای محرمانگی وجودندارد. هنوز ناسازگاری‌هایی در باب این مسئله وجوددارد که باید برطرف‌شوند. به موازات پیشرفت‌های داده‌کاوی، سوالات متنوعی افزایش می‌یابند شامل اینکه نهادهای شهری و دولتی تا چه اندازه می‌بایست داده‌های تجاری را با داده‌های دولتی استفاده و ترکیب‌کنند، آیا منابع داده به منظورهایی غیر از هدف اصلی طراحی می‌شوند و کاربردهای ممکن از اعمال محرمانگی چیست و غیره .

نتیجه‌گیری

در این مقاله یک مطالعه برای داده‌کاوی در شهر الکترونیک ارائه شد. این مطالعه شامل ۵ بخش بود. کاربردهای مهم داده‌کاوی و چالش‌های آن در پلت‌فرم شهر الکترونیک توصیف شد که کمک تحلیلی برای تصمیم‌گیری، نظارت و یا بازنگری می‌باشد. این مطالعه می‌تواند منافعی را در اختیار سهام‌داران مختلف و صاحبان اختیار که نیاز به دستگیری دانش مخفی و ضمنی از شهروندان، سازمان‌ها و یا کسب و کارها دارند، قرار دهد.

شهر الکترونیک، در قلمرو قدرت خود بزرگ‌ترین گردآورنده داده است. داده‌کاوی یک ابزار بالقوه سودمند برای کاوش داده‌های دولت و تولید پشتیبانی خوب برای تصمیمات و تحلیل‌های دولت می‌باشد. و ازاین‌رو دولت به منظور افزایش رضایت شهروندان و کسب‌وکارها و افزایش کارایی و بهره‌وری اقتصادی، بایستی تصمیم‌گیری‌ها و سامان‌بخشیدن به استراتژی‌ها و تاکتیک‌ها را به‌طور پویا با استفاده از کاوش داده‌های استخراج‌شده انجام دهد. و نیز بستر لازم را جهت پشتیبانی پیوسته، پذیرش توسط افراد و ایجاد آموزش‌های لازم فراهم‌سازد.

در حالی که توانایی‌های تکنولوژیکی مهم هستند مسائل و چالش‌هایی در پیاده‌سازی و نیز موفقیت داده‌کاوی در شهر الکترونیک وجوددارد که بایستی به آنها توجه شده و آنها را برطرف‌کرد. فاکتورهایی وجوددارند که در این تحقیق به‌اختصار به چند فاکتور از قبیل مقیاس‌پذیری و عدم پشتیبانی همزمان، محرمانگی داده‌ها، دقت نتایج حاصل و غیره اشاره‌شد. بنابراین جوامع شهری، سیاسی و دولتی بایستی از مسائل و چالش‌هایی که در رابطه با پیاده‌سازی و کاربردهای داده‌کاوی وجوددارد آگاه‌باشند. در مطالعات آینده به پیشنهاد روش‌ها و استراژی‌هایی جهت حل مشکلات داده‌کاوی در شهر الکترونیک خواهیم‌پرداخت.

از طرفی اکنون بیشتر از ۸۰ درصد از دانش سازمان‌ها به صورت متن ذخیره‌می‌شود و انتظار می‌رود که متن‌کاوی یکی از مهم‌ترین تکنولوزی‌ها در آینده باشد، بنابراین ایجاد یک چهارچوب برای استفاده از متن‌کاوی در سیستم‌های شهر الکترونیک و بررسی مشکلات و چالش‌های آن، زمینه تحقیقات بعدی را در این حوزه تشکیل می‌دهد. استفاده از داده‌کاوی مبتنی بر آنتولوژی برای یکپارچگی بیشتر سرویس-های شهری و رسیدن به افزایش کارایی و قابلیت اطمینان در شهر الکترونیک نیز می‌تواند زمینه‌ای برای تحقیقات آتی باشد.

مراجع و ماخذ

مراجع و ماخذ فارسی

۱. دانشنامه آزاد ویکی پدیا

۲. ماهنامه عملی آموزشی تدبیر شماره ۱۵۶

۳. مهریزی، حائری، علی اصغر ، «داده‌کاوی: مفاهیم، روش‌ها و کاربردها» (۱۳۸۲) پایان‌نامه کارشناسی ارشد آمار اقتصادی و اجتماعی، دانشکده اقتصاد، دانشگاه علامه طباطبائی.

۴. زعفریان، رضا و زعفریان،قاسم، «مروری بر داده‌کاوی» (۱۳۸۰) فصلنامه صنایع، شماره ۲۹

۵. شاه‌سمندی، پرستو «داده‌کاوی در مدیریت ارتباط با مشتری» (۱۳۸۴)، مجله تدبیر شماره ۱۵۶.

۶. گودرزی، حمیدرضا، مترجم «داده‌کاوی چیست»، نشریه گزیده مطالب آماری، مرکز آمار ایران، شماره ۵۲.

۷. جمالی، آرمان - شهر الکترونیکی، بستر ورود به رقابت‌های عصر سیبرنتیک *(http/www.editorial.com)*

مراجع و ماخذ لاتین و سایتهای اینترنتی

8. *Barbara Mento and Brendan Rapple, SPEC Kit 274: Data mining and data warehousing, Association of Research Libraries, Washington, DC (2003, July)*
9. http://www.infotechera.com/
10. http://www.ece.ut.ac.ir/dbrg/index.htm
11. *http://www.irandoc.ac.ir/index.htm*
12. *http://www.arts.uci.edu/dobrain/gems.980415b.htm*
13. Two Crows Corporation. *Introduction to data mining and knowledge discovery*, (3rd ed.) (p. 4). Potomac, MD7 Two Crows Corporation, 1999.
14. Rahman Hakikur , *Social and Political Implications of Data Mining: Knowledge Management in E-Government*, Published in the United States of America byInformation Science Reference (an imprint of IGI Global), Hershey • New York, ISBN 978-1-60566-230-5, 2009.
15. Kazienko Przemyslaw, Adamski Michal., *"AdROSA—Adaptive Personalization of Web Advertising"*, *Information Sciences* (177) ,2269 – 2295(2007).
16. Zhou Ping, Le Zhongjian, *A Framework for Web Usage Mining in Electronic Government*, Integration and Innovation Orient to E-Society, Volume 2, pp. 487-496, in IFIP International Federation for Information Processing, December 27,2007.
17. Sabol Tomas, Mach Marian, *Semantic Web in e-Government*, http://www.accessegov.org/acegov/uploadedFiles/webfiles/cffile_10_9_06_10_02_56_AM.pdf
18. http://www.dbmsmag.com, 1998

19. Piatetsky-Shapiro Gregory, Djeraba Chabane, Getoor Lise, *What are the grand challenges for data mining?: KDD-2006 panel report,* SIGKDD Explorations Newsletter archive, Volume 8 , Issue 2, Pages 70 – 77, December 2006, http://portal.acm.org/citation.cfm?id=1233321.1233330
20. http://www.sqliran.com/SQLIran/Mod_Core/Pages/Services/DataMining.aspx
21. http://www.iranwbs.com
22. http://www.bedkaco.com
23. http://www.articles.ir
24. Hand. D.J (1998): "Review of Data mining", The American statistician, 52, 112-118.
25. samashirwan.wordpress.com
26. Other Internet Site…