داده کای در مدیریت ارتباط با مشتری

هومن حری

دانشجوی مهندسی فن آوری اطلاعات

horrihooman@gmail.com

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 پردازش کرد و اطلاعات حاصل از این پردازش را در اختیار کاربران قرار داد.

با استفاده ار 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 چقدرحرفه ای 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 مالی بسیار بالا است.

بنابراین می شود گفت که درحال حاضر یک تغییر الگو از مدل سازی و تحلیل های کلاسیک برپایه اصول اولیه به مدل های درحال پیشرفت و تحلیل های مربوط بطور مستقیم از داده ها وجود دارد.

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 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 شودبنابراین می توان گفت در داده کاوی تئوریهای پایگاه داده ها، هوش مصنوعی، یادگیری ماشین وعلم 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 گیگابایت یا ترابایت، مواجه باشیم که از این نظر یکی از بزرگترین بازارهای هدف، انبارجامع داده ها، مراکز داده وسیستم های پشتیبانی تصمیم برای بدست آوردن تخصص هایی در صنایعی مثل شبکه های توزیع مویرگی، تولیدف مخابرات، بیمه و ...می باشد.

مدیریت ارتباط با مشتری به همه فرآیندها و فناوری‌هایی گفته می‌شود که در شرکت‌ها و سازمان‌ها برای شناسایی، ترغیب، گسترش، حفظ و ارائه خدمت به مشتریان به کار می‌رود.

سازمان‌ها با استفاده از CRM می توانند چرخه فروش را کوتاهتر و وفاداری مشتری به ایجاد روابط نزدیکتر و درآمد را افزایش دهند. سیستم مدیریت روابط با مشتری می‌تواند کمک کند تا مشتریان موجود حفظ شوند و مشتریان جدیدی جذب شوند. سازمان‌ها برخی روش‌هایی را شامل مدیریت ارتباط با مشتری،

تحلیل ارزش مشتری، استراتژی سازمانی و ساز و کارهای خدماتی که کارایی ارتباطات مشتری را بهبود می‌دهد بکار می برند. مدیریت ارتباط با مشتری استراتژی ای برای کسب مشتریان جدید و نگهداشتن آنها است. مدیریت ارتباط با مشتری عملیاتی شامل تمام فعالیتهای مرتبط با مشتریان بی واسطه همچون شرکتها می باشد.

ارتباط بین این دو(داده کاوی و مدیریت ارتباط با مشتری) خیلی مهم و اساسی می باشد زیرا شرکت ها با داده کاوی خواسته های مشتریان خود را می شناسند و بر اساس با آن به مشتریان خود سرویس می دهند

در این مقاله در بخش اول با مقدمه ای از داده کاوی و تاریخچه ای در این مورد را بررسی می کنیم و در بخش بعد با داده کاوی اشنا می شویم بعد با وب کاوی که اساس بسیاری از شرکت ها برای استخراج داده های خام از پرتال های خود هستند اشنا می شویم در بخش چهارم با مدیریت ارتباط با مشتری اشنا می شویم و در فصل آخر ارتباط بین این دو را برسی می کنیم

فصل اول

مقدمه

۱-۱ مقدمه :

از هنگامی که رایانه در تحلیل و ذخیره سازی داده ها بکار رفت (1950) پس از حدود 20 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 شده و همچنین تعداد پایگاه داده ها با سرعت بیشتری رشد نمود. این در حالی است که تعداد متخصصین تحلیل داده ها با این سرعت رشد نکرد. حتی اگر چنین امری اتفاق می افتاد، بسیاری از پایگاه داده ها چنان گسترش یافته‌اند که شامل چندصد میلیون یا چندصد میلیارد رکورد ثبت شده هستند.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 عظیمی از داده ها شده است.

این واقعیت، 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، شرکت و غیره برای امور بازرگانی، پرسنلی، آموزشی، آماری و غیره پایگاه داده ها ایجاد یا خریداری شده است. به طوری که این پایگاه داده ها برای مدیران، برنامه ریزان، پژوهشگران جهت، تصمیم گیری های راهبردی، تهیه گزارش های مختلف، توصیف وضعیت جاری خود و سایر اهداف می تواند مفید باشد. بسیاری از این داده ها از نرم افزارهای تجاری، مثل کاربردهای مالی، ERPها، CRMها و web log ها، می آیند. نتیجه این جمع آوری داده ها این می‌شود که در سازمانها، داده ها غنی ولی دانش ضعیف، است. جمع آوری داده ها، بسیار انبوه می‌شود و بسرعت اندازه آن افزایش می یابد و استفاده عملی از داده ها را محدود می سازد.

داده‌کاوی استخراج و تحلیل مقدار زیادی داده بمنظور کشف قوانین و الگوهای معنی دار در آنهاست. هدف اصلی داده کاوی، استخراج الگوهایی از داده ها، افزایش ارزش اصلی آنها و انتقال داده ها بصورت دانش است.

داده‌کاوی، بهمراه OLAP، گزارشگری تشکیلات اقتصادی(Enterprise reporting) و ETL، یک عضو کلیدی در خانواده محصول (Business Intelligence(BI، است. حوزه‌های مختلفی وجود

دارد کـه در آنهـا حجـم بسـیاری از داده در پایگـاه‌داده‌هـای متمرکـز یـا توزیـع شـده ذخیـره مـی‌شـود. برخی از آنها به قرار زیر هستند: **Error! Reference source not found.**

1- کتابخانه دیجیتـال: یـک مجموعـه سـازماندهی شـده از اطلاعـات دیجیتـال کـه بصـورت مـتن در پایگاه‌داده‌های بزرگی ذخیره می شوند.
2- آرشیو تصویر: شامل پایگاه‌داده بزرگی از تصاویر به شکل خام یا فشرده.
3- اطلاعـات زیسـتی: بـدن هـر انسـانی از 50 تـا 100 هـزار نـوع ژن یـا پـروتئین مختلـف سـاخته شـده اسـت. اطلاعـات زیسـتی شـامل تحلیـل و تفسـیر ایـن حجـم عظـیم داده ذخیـره شـده در پایگاه‌داده بزرگی از ژنهاست.
4- تصاویر پزشکی: روزانـه حجـم وسـیعی از داده‌هـای پزشـکی بـه شـکل تصـاویر دیجیتـال تولیـد مـی‌شـوند، ماننـد EKG، MRI، ACT، SCAN و غیـره. اینهـا در پایگـاه‌داده‌هـای بزرگـی در سیستم‌های مدیریت پزشکی ذخیره می شوند.
5- مراقبـت‌هـای پزشـکی: بجـز اطلاعـات بـالا، یکسـری اطلاعـات پزشـکی دیگـری نیـز روزانـه ذخیـره مـی‌شـود ماننـد سـوابق پزشـکی بیمـاران، اطلاعـات بیمـه درمـانی، اطلاعـات بیمـاران خاص و غیره.
6- اطلاعـات مـالی و سـرمایه‌گـذاری: ایـن اطلاعـات دامنـه بزرگـی از داده‌هـا هسـتند کـه بـرای داده‌کاوی بسیار مطلوب می‌باشـند. از ایـن قبیـل داده‌هـا مـی‌تـوان از داده‌هـای مربـوط بـه سـهام، امـور بـانکی، اطلاعـات وام‌هـا، کـارت‌هـای اعتبـاری، اطلاعـات کـارت‌هـای ATM، و کشـف کلاه‌برداری‌ها می باشد.
7- ساخت و تولید: حجـم زیـادی از ایـن داده‌هـا روزانـه بـه اشـکال مختلفـی در کارخانـه‌هـا تولیـد مـی‌شـود. ذخیـره و دسترسـی کـارا بـه ایـن داده‌هـا و تحلیـل آنهـا بـرای صـنعت تولیـد بسـیار بااهمیت است.
8- کسـب و کـار و بازاریـابی: داده لازم اسـت بـرای پـیش‌بینـی فـروش، طراحـی کسـب و کـار، رفتار بازرایابی، و غیره.
9- شـبکه راه‌دور: انـواع مختلفـی از داده‌هـا در ایـن صـنعت تولیـد و ذخیـره مـی شـوند. آنهـا بـرای تحلیل الگوهای مکالمات، دنبـال کـردن تمـاس‌هـا، مـدیریت شـبکه، کنتـرل تـراکم، کنتـرل خطـا و غیره، استفاده می‌شوند.
10- حـوزه علـوم: ایـن حـوزه شـامل مشـاهدات نجـومی، داده زیسـتی، داده ژنومیـک، و غیـره است.

11- WWW: یک حجم وسیع از انواع مختلف داده که در هر جایی از اینترنت پخش شده‌اند.

در بیشتر این حوزه‌ها، تحلیل داده‌ها یک روال دستی بود. یک تحلیلگر کسی بود که با داده‌ها بسیار آشنا بود و با کمک روش‌های آماری، خلاصه‌هایی تهیه و گزارشاتی را تولید می‌کرد. در یک حالت پیشرفته‌تر، از یک پردازنده پیچیده پرسش استفاده می‌شد. اما این روش‌ها با افزایش حجم داده‌ها کاملا بلااستفاده شدند.

واژه های «داده‌کاوی» و «کشف دانش در پایگاه داده»اغلب به صورت مترادف یکدیگر مورد استفاده قرار می گیرند. کشف دانش به عنوان یک فرآیند در شکل1 نشان داده شده است.

کشف دانش در پایگاه داده فرایند شناسایی درست، ساده، مفید، و نهایتا الگوها و مدلهای قابل فهم در داده ها می‌باشد. داده‌کاوی، مرحله‌ای از فرایند کشف دانش می‌باشد و شامل الگوریتمهای مخصوص داده‌کاوی است، بطوریکه، تحت محدودیتهای مؤثر محاسباتی قابل قبول، الگوها و یا مدلها را در داده کشف می کندبه بیان ساده‌تر، داده‌کاوی به فرایند استخراج دانش ناشناخته، درست، و بالقوه مفید از داده اطلاق می‌شود. تعریف دیگر اینست که، داده‌کاوی گونه‌ای از تکنیکها برای شناسایی اطلاعات و یا دانش تصمیم‌گیری از قطعات داده می‌باشد، به نحوی که با استخراج آنها، در حوزه‌های تصمیم‌گیری، پیش بینی، پیشگویی، و تخمین مورد استفاده قرار گیرند. داده‌ها اغلب حجیم، اما بدون ارزش می‌باشند، داده به تنهایی قابل استفاده نیست، بلکه دانش نهفته در داده ها قابل استفاده می باشد. به این دلیل اغلب به داده کاوی، تحلیل داده ای ثانویه گفته می‌شود.

استخراج دانش در پایگاه‌داده (KDD)، بعنوان روالی برای شناسایی الگوهای معتبر، جدید، بالقوه مفید، و سرانجام قابل فهم در داده‌ها، تعریف شده است. روال سراسری شامل تبدیل داده سطح-پایین به دانش سطح-بالاست. این روال یک روال تعاملی و تکراری است که شامل مراحل زیر می‌باشد. **Error! Reference source not found.**

1- درک دامنه کاربرد: این شامل دانش قبلی مرتبط و اهداف کاربرد است.
2- استخراج مجموعه داده هدف: این چیزی نست جز انتخاب یک مجموعه داده یا یک زیرمجموعه از متغیرها، با استفاده از تکنیک‌های رتبه‌بندی و انتخاب است.
3- پیش پردازش داده: این مرحله برای افزایش کیفیت داده بکار گرفته شده برای داده‌کاوی، لازم است. همچنین برای بهبود کارایی کاوش داده لازم است. پیش پردازش داده شامل پاکسازی داده، انتقال داده، یکپارچه سازی داده، کاهش یا فشرده‌سازی داده برای نمایش فشرده، و غیره است.
4- داده‌کاوی: این مرحله شمل اعمال یکی از الگوریتم‌های داده‌کاوی است.

5- تفسیر: شامل تفسیر الگوهای استخراج شده، و تا حد امکان، بصری سازی این الگوهاست. بصری سازی یک کمک‌کننده مهم در قابل فهم سازی الگوهاست.
6- استفاده از دانش استخراج شده: این مرحله شامل تلفیق این دانش با کارایی سیستم و گرفتن تصمیمات عملی براساس این دانش است.

بیشتر تکنیکهای داده‌کاوی حداقل به عنوان الگوریتمهای آکادمیک از سالها یا دهه های قبل وجود داشته اند. تنها در دهه اخیر است که داده‌کاوی تجاری نقش عمده ای را بازی کرده است.

چرا امروزه ما به داده‌کاوی گرایش داریم؟ در زیر تعدادی از دلایل آن آورده شده: **Error! Reference source not found.**

1. مقدار زیاد داده در دسترس: در دهه اخیر، قیمت سخت افزار بویژه فضای دیسک سخت، بسیار کاهش یافته است. و به دنبال آن، تشکیلات اقتصادی مقدار زیادی از داده ها را از کاربردهای زیادی گردآوری کرد. با این انفجار داده ها، تشکیلات اقتصادی می خواهند که الگوهای پنهان در این داده ها را برای هدایت استراتژی های تجارت خود بکار گیرند. داده‌کاوی هنگامی بیشترین معنی را پیدا می کند که داده های زیادی وجود داشته باشد. اغلب الگوریتم های داده‌کاوی نیازمند میزان زیادی از داده ها هستند تا مدلهایی را ترتیب دهند که بعداً برای دسته بندی، تخمین، پیش بینی یا سایر کارکردهای داده‌کاوی مورد استفاده قرار گیرند.
2. افزایش رقابت: رقابت بعلت وجود بازارهای مدرن و کانالهای توزیع مثل اینترنت و ارتباطات راه دور، بطور فزاینده ای در حال افزایش است. تشکیلات اقتصادی با رقابتهای جهان وب مواجه اند و کلید موفقیت در تجارت، حفظ مشتریان کنونی و بدست آوردن مشتریان جدید است. داده کاوی، تکنولوژی هایی دارد که اجازه می دهد که تشکیلات تجاری فاکتورهایی را برای مواجه با این زمینه ها تحلیل کند.
3. آماده بودن تکنولوژی آن: داده‌کاوی قبلا فقط در حوزه آکادمیک قرار داشت، اما در حال حاضر بسیاری از این تکنولوژی ها کامل شده اند و برای اعمال در صنعت آماده اند. الگوریتم ها، بسیار دقیق تر و کاراتر شده اند و می توانند بطور فزاینده ای داده های پیچیده را مدیریت کنند. بعلاوه رابط برنامه نویسی کاربردهای داده کاوی(APIها)، اکنون استاندارد شده اند، که به توسعه دهندگان این امکان را می دهند که کاربردهای داده‌کاوی بهتری بسازند.

4. علاقه به مدیریت روابط با مشتریان فراوان است: در طیف وسیعی از صنایع، شرکتها به این بینش رسیده اند که مشتریان برای سازمان حیاتی هستند. و اطلاعات درباره آن مشتریها یکی از دارایی های اساسی سازمان می باشد. اطلاعاتی که شرکتها درباره مشتریانشان دارند نه تنها برای خودشان بلکه برای دیگران هم ارزشمند است. اطلاعات یک محصول است. یک شرکت کارت اعتباری چیزهایی می داند که شرکتهای خطوط هوایی دوست دارند بدانند یعنی چه کسی بلیطهای پرواز متعددی می خرد. گوگل می داند مردم در وب دنبال چه چیزی هستند و از این شناخت با فروش لینکهایی با پشتیبان مالی بهره میبرد. در واقع هر شرکتی که داده های با ارزش جمع آوری می کند در موقعیت یک واسطه اطلاعات قرار دارد.

۲-۱ تاریخچه

با رشد فناوری اطلاعات و روش های تولید و جمع آوری داده ها ، پایگاه داده های مربوط به داده های تبادلات تجاری ، کشاورزی ، اینترنت ، جزییات مکالمانی تلفنی ، داده های پزشکی و ... سریع تر از هر روز جمع آوری و انبارش میشود . لذا از اواخر دهه ی 80 میلادی بشر به فکر دستیابی به اطلاعات نهفته در این پایگاه های داده های حجیم افتاد . زیرا سیستم های سنتی قادر به این کار نبودند . به دلیل رقابت در عرصه های سیاسی ، نظامی ، اقتصادی ، علمی و اهمیت دست یابی به اطلاعات در کم ترین زمان بدون دخالت انسان علم تجزیه و تحلیل داده ها یا داده کاوی پا به عرصه گذاشت .

داده کاوی فر آیندی است که در آغاز دهه 90 مطرح شد و با نگرشی نو به مساله استخراج اطلاعات از پایگاه داده ها میپردازد .از سال 1995 داده کاوی به صورت جدی وارد مباحث آمار شد و در سال 1996 اولین شماره مجله کشف دانش و معرفت از پایگاه داده ها منتشر شد . محقق های نظیر براچمن و آناند کلیه مراحل واقع گرایانه و رو به جلو کشف دانش از پایگاه داده ها را تشخیص دادند .

در حاضر ، داده کاوی مهمترین فناوری جهت بهره برداری مواثر از داده های حجیم است و اهمیت آن رو به فزونی است . به طوری که تخمین زده شده است که مقدار داده ها در جهان در هر 20 ماه حدود دو برابر میشود . در یک تحقیق که بر روی گروه های تجاری بسیار بزرگ در جمع آوری داده ها صورت گرفت مشخص گرفت که 19 درصد از این گروه ها دارای پایگاه داده هایی با سطح بیشتر از 50 گیگابایت میباشند و 59 در صد آن ها انتظار دارند که در آینده نزدیک در چنین سطحی قرار گیرند.

در صنایعی مانند کارت های اعتباری و ارتباطات و فروشگا های زنجیره ای و خرید های الکترنیکی و اسکنر های بار کد خوان هر روزه داده های زیادی تولید و ذخیره میشوند . افزایش سرعت کامپیوتر باعث بوجود آمدن الگوریتم هایی شده است که قدرت تجزیه و تحلیل های بیشتری دارد بدون اینکه محدودیت در زمینه ظرفیت و سرعت کامپیوتر داشته باشد .

در سال 1989 و 1991 کارگاه های کشف دانش و معرفت از پایگاه داده ها توسط پیاتتسکی و همکارانش برگذار شد . در فواصل سال های 1991 تا 1994 کارگاه های کشف دانش و معرفت از پایگاه داده ی توسط فییاد و پیاتتسکی و دیگران برگزار شد . به طور رسمی اصطلاح داده کاوی برای اولین بار توسط فییاد در اولین کنفرانس بین المللی " کشف دانش و داده کاوی " در سال 1995 مطرح شد . امروزه کنفرانس های مختلفی در این زمینه در سراسر دنیا برگزار شد. افزایش داده های بسیار باعث پیدایش فرصت های تازه برای کار در علوم مهندسی و کسب و کار شده است . زمینه داده کاوی و کشف دانش از پایگاه داده ها به عنوان یک رشته علمی جدید در مهندسی علوم کامپیوتر ظهور کرده است . مهندسی صنایع با حوزه های گوناگون و در برداشتن فرصت های بی نظیر اکنون برای کاربرد داده کاوی و کشف دانش از پایگاه داده ها و برای توسعه مفاهیم و روش های تازه در این زمینه آماده ه است. فر آیند های صنعتی زیادی اکنون برای مطمین شدن از کیفیت سفارشات محصول و کاهش هزینه های محصول به طور خودکار و کامپیوتری شده اند.

فصل دوم

داده کاوی

۱-۲ تعاریف داده کاوی

در متون آکادمیک تعاریف گوناگونی برای داده کاوی ارائه شده اند. 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 :

1- 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 .

2- فرایند نیم خودکار تجزیه و تحلیل پایگاه داده های بزرگ به منظور یافتن الگوهای مفید اطلاق می شود.
3- داده کاوی یعنی فرایند جستجو در یک پایگاه داده ها برای یافتن الگوهایی میان داده ها .
4- داده کاوی یعنی تجزیه و تحلیل مجموعه داده های قابل مشاهده برای یافتن روابط مطمئن بین داده ها.
5- داده کاوی یعنی استخراج دانش کلان ، قابل استناد و جدید از پایگاه داده ها ی بزرگ.

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.

داده کاوی تحلیل اکتشافی داده درمیان چیزهای دیگرنامیده می شود . حجم داده های تولید شده ازثبت پول ، اسکن ، پایگاه داده های خاص موضوعی کل شرکت کشف و تحلیل شده ، کاهش یافته ودوباره استفاده می شوند . تحقیقات درمیان مدل های مختلف پیشنهاد شده برای پیشگویی فروش ، واکنش بازار و سود انجام می شوند . روش های اماری کلاسیک ، اصول داده کاوی می باشند . متد های AI خود کارنیزاستفاده می شوند . به علاوه شناسایی اصولی به واسطه متدهای اماری کلاسیک هنوز پایه و اساس داده کاوی هستند . بعضی از ابزارهای توسعه توسط زمینه

تحلیل اماری ازطریق کنترل خودکار(همراه با بعضی راهنمایی های کلیدی بشر) در برخورد با داده ها مهار شده اند .

داده کاوی عبارت است ازفرایند خودکار کشف دانش و اطلاعات از پایگاه‌های داده‌ای .این فرایند تکنیکهایی ازهوش مصنوعی را برروی مقادیرزیادی داده اعمال می کند تا روندها ، الگوها و روابط مخفی را کشف کند . ابزارهای داده کاوی برای کشف دانش یا اطلاعات از داده‌ها به کاربر اتکا نمی کنند ، بلکه فرایند پیشگویی واقعیت‌ها را خودکار می سازند.این تکنولوژی نوظهور،اخیراً به طور فزاینده‌ای درتحلیل‌ها مورد استفاده قرار می‌گیرد .

۲-۱-۱ تعریف جامع:

"داده کاوی فرآیندی است که طی آن با استفاده از ابزار های 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."

در تعاریفی که از داده کاوی ارائه شد به اصطلاح "فرایند" اشاره شد. حتی در بعضی محیط های حرفه ای این نظر وجود دارد که داده کاوی شامل انتخاب و بکارگیری ابزارهای مبتنی بر کامپیوتر برای حل مسائل فعلی و بدست آوردن یک راه حل بطور اتوماتیک و خودکار میباشد.

برای اموزش داده کاوی، باید بر مفاهیم و روش های اعمال شده برخلاف همه جاذبه های ابزارهای مبتنی بر کامپیوتر که امور رابا جزئیات ودستورات با فرمت های خاصی باید به خیلی از سوالات از جمله چگونگی طراحی واستفاده از فرایندها را پاسخ داد به جای بیان جزئیات عملی ابزار مختلف داده کاوی تکیه نمود

2-1-2 مراکز داده چیست

مراگز داده یا سرور فارم ، ساختمان های خاصی هستند که در آن ها انبوهی از تجهیزات آی تی نظیر سیستم های کامپیوتر و مولفه های مرتبط با آن نسب شده تا پیش نیاز های لازم برای عملکرد درست و بی نقض سایت های سازمان فراهم شود . این مرکز همچنین وظیفه ی برقراری نظم و امنیت سایت و تهیه بستر ارتباط سازمان را بر عهده دارد .

3-1-2 برخی از کاربردهای داده کاوی در محیطهای واقعی عبارتند از :

1- خرده فروشی
2- تعیین الگوهای خرید مشتریان
3- تجزیه و تحلیل سبد خرید بازار
4- پیشگویی میزان خرید مشتریان از طریق پست(فروش الکترونیکی)
5- بانکداری
6- پیش بینی الگوهای کلاهبرداری از طریق کارتهای اعتباری
7- تشخیص مشتریان ثابت
8.- تعیین میزان استفاده از کارتهای اعتباری بر اساس گروههای اجتماعی
9- بیمه
10- تجزیه و تحلیل دعاوی
11- یشگویی میزان خرید بیمه نامه های جدید توسط مشتریان
12- پزشکی
13- تعیین نوع رفتار با بیماران و پیشگویی میزان موفقیت اعمال جراحی
14- تعیین میزان موفقیت روشهای درمانی در برخورد با بیماریهای سخت

4-1-2 مسائل کسب و کار برای داده کاوی.**Error! Reference source not found. :**

تکنیکهای داده‌کاوی می تواند برای کاربردهای بسیاری بکار رود، در زیر تعدادی از مسائل کسب و کار معمولی را که می‌توان با داده‌کاوی برای آنها پاسخی یافت، شرح داده می‌شود:

1. تحلیل رویگردانی: کدام مشتریان بیشتر تمایل دارند بسمت رقیب ما بروند؟ امروزه صنایع تله کام، بانکداری و بیمه، بیش از دیگران در معرض این رقابت ها هستند. بطور متوسط، هر مشترک جدید تلفن همراه، برای شرکت تلفن، هزینه ای بالغ بر 200 دلار در بازار سرمایه گذاری دارد. هر کسب و کاری می خواهد مشتریان بیشتری را کسب کند. تحلیل رویگردانی می تواند به مدیران بازاریابی کمک کند تا دلایل رویگردانی مشتری را درک کند و روابط مشتری را بهبود دهد و وفاداری مشتری را افزایش دهد.
2. فروش متقاطع: مشتریان چه محصولاتی را دوست دارند بخرند؟ فروش متقاطع برای خرده فروشان یک چالش تجاری بزرگ است. بسیاری از خرده فروشان، بویژه خره فروشان online، برای افزایش فروش خود از این ویژگی استفاده می کنند. برای مثال، اگر شما به یک کتابفروشی Online مثل Amazon.com برای خرید یک کتاب، بروید، شما متوجه شده اید که وب سایت مجموعه ای از پیشنهادات درباره کتابهای مرتبط را به شما پیشنهاد می کند. این پیشنهادات را می توان از تحلیل دده کاوی گرفت.
3. کشف تقلب: آیا این ادعای بیمه، کلاهبرداری است؟ شرکت های بیمه، روزانه هزاران دعوی بیمه دارند. برای آنها مهم است که درباره هر مورد تحقیق و بررسی کنند. داده‌کاوی می تواند برای شناسایی دعاوی ای که بیشتر نادرست هستند، کمک کند.
4. مدیریت ریسک: آیا باید این درخواست وام برای این مشتری تصویب شود؟ این یک سوال بسیار رایج در سناریوهای بانکی است. تکنیک های داده‌کاوی می توانند برای رتبه بندی سطح ریسک یک مشتری، بکار روند، و به مدیر در گرفتن یک تصمیم مناسب برای هر کاربرد، کمک کنند.
5. قطعه‌بندی مشتریان: چه کسی مشتری من است؟ قطعه‌بندی مشتریان به مدیران بازاریابی کمک می‌کند که تفاوت های پروفایل‌های مشتریان را درک کنند و عمل بازاری مناسبی را بر مبنای هر بخش، انجام دهند.
6. تبلیغات هدف‌دار: چه بنر تبلیغی باید برای یک بازدیدکننده خاص، نمایش داده شود؟ فروشندگان وب و سایت های پورتال تمایل دارند که محتوای خود را برای مشتریانشان سفارشی کنند. با استفاده از الگوهای ناوبری مشتری یا خرید online، این سایت ها می توانند راه حل های داده‌کاوی را برای نمایش تبلیغات هدف دار برای مشتریانشان، بکار برند.
7. پیش‌بینی فروش: من چه نمونه‌هایی را در این فروشگاه در این هفته خواهم فروخت؟ تکنیک‌های پیش‌بینی داده‌کاوی می تواند برای پاسخ به انواع این پرسش های مرتبط با زمان، بکار روند.

۲-۲ چرخه تعالی داده کاوی چیست؟

باید بتوان داده‌ها را به اطلاعات، اطلاعات را به عمل و عمل را به ارزش تبدیل کرد. این را در یک کلام چرخه تعالی داده‌کاوی می‌نامند. به منظور دستیابی به این هدف لازم است داده‌کاوی به صورت یک فرآیند ضروری در کنار سایر فرآیندها نظیر بازاریابی، فروش، پشتیبانی مشتری وکنترل موجودی درآید.

در ادبیات بازاریابی داده‌کاوی را بسیار آسان جلوه می‌دهند بطوریکه تنها کافیست الگوریتم‌های خودکار تهیه شده توسط بهترین دانشمندان دانشگاهی نظیر شبکه‌های عصبی، درخت‌های تصمیم‌گیری و الگوریتم ژنتیک را به کار برد تا در مسیر موفقیت قرار گیرید. اگرچه این الگوریتم ها مهمند، اما راه حل های داده کاوی چیزی فراتر از مجموعه‌ای از تکنیکها و ساختارهای داده‌ای قوی است. این تکنیکها را باید در جای مناسب و با داده‌های صحیح بکار برد. چرخه تعالی داده‌کاوی یک فرآیند یادگیری تکراری و مرحله‌ای است که بر اساس نتایج در طول زمان تهیه می‌گردد. موفقیت در بکارگیری داده‌ها، وضعیت یک سازمان را از واکنشی به کنشگرا تبدیل خواهد نمود. با استفاده از چرخه تعالی داده‌کاوی مطرح شده در این کارگاه، بیشترین سود از تکنیکهای داده‌کاوی به دست می‌آید.

این چرخه شامل چهار مرحله است:

1. تشخیص مشکل کسب وکار و تجارت
2. واکاوی داده‌ها برای تبدیل آن‌ها به اطلاعات عملی
3. کار روی اطلاعات
4. بررسی نتایج

کلید موفقیت، در گنجاندن داده کاوی در فرآیندهای تجاری است و اینکه بتوان راههای ارتباطی میان داده کاوان تکنیکی و کاربران تجاری نتایج یافت.

1- تعیین فرصت‌های کسب‌وکار و تجارت :

چرخه تعالی داده کاوی با تعیین فرصت‌های واقعی کسب‌وکار و تجارت آغاز می گردد. متأسفانه تحلیل‌گران زیادی هستند که معمولاً تلاش‌هایشان بی ثمر می ماند زیرا آنها مسائلی را حل می‌کنند که به

کسب‌وکار و تجارت کمکی نمی‌نماید. داده‌کاوان خوب می‌خواهند این وضعیت برایشان پیش نیاید. اجتناب از تلاشهای تحلیلی بی‌حاصل با اراده برای کار روی نتایج آغاز می‌گردد.

بیشتر فرایندهای تجاری معمول کاندیداهای خوبی برای داده کاوی هستند:

1- برنامه‌ریزی برای معرفی محصول جدید
2- برنامه‌ریزی برای بازاریابی مستقیم
3- فهم احساسات و گلایه‌های مشتریان

اگر نتوان نتایج داده‌کاوی را سنجید نمی‌توان از آن چیزی یاد گرفت و آنگاه هیچ چرخه تعالی وجود نخواهد داشت. سنجش تلاش‌های صورت‌گرفته در گذشته سوالاتی را درباره تجارت، فرصت‌های جدیدی از داده کاوی را فراهم می‌کند:

1- چه نوع مشتریانی به جدید ترین اقدام پاسخ داده اند؟
2- بهترین مشتریان کجا زندگی می‌کنند؟
3- آیا صرف زمان طولانی در پای دستگاههای خودپرداز دلیل گلایه آنهاست؟
4- آیا مشتریان سودآور از خدمات پشتیبانی مشتریان استفاده می‌کنند؟
5- چه نوع محصولات دیگری را باید با یک محصول خاص تبلیغ نمود؟

مصاحبه با خبرگان تجارت، روش خوب دیگری برای آغاز است؛ زیرا ممکن است کسانی که در داخل تجارت هستند، با داده‌کاوی آشنا نباشند. آنها ممکن است ندانند چگونه روی نتایج کار کنند.

با بیان ارزش داده‌کاوی برای یک سازمان، چنین مصاحبه‌هایی امکان ایجاد ارتباطی دوسویه را فراهم می‌کند

2- مشکلات در راه کسب توانایی استفاده از نتایج داده کاوی :

داده‌ها در اشکال گوناگون و از سیستم‌های چندگانه‌ای ظاهر می‌شوند. تشخیص منابع درست داده‌ها و گرد آوردن آنها کنار یکدیگر از عوامل حیاتی در موفقیت می‌باشد.

هر پروژه داده‌کاوی مسائل خاص خود در ارتباط با داده ها را دارد: سیستم‌های اطلاعاتی ناهماهنگ، اطلاعاتی که هر چند ماه یکبار باز نویسی می‌شوند. در اینجاست که الگوریتم‌های داده‌کاوی بکار می‌آیند.

3- انجام عمل :

انجام عمل، هدف چرخه متعالی داده کاوی است. همانطور که ذکر شد، عمل می تواند اشکال مختلفی داشته باشد. داده‌کاوی سبب می شود تصمیمات کسب و کار آگاهانه‌تر اتخاذ شود. و در طول زمان، تصمیمات با آگاهی بیشتر، نتایج بهتری بدنبال دارد.

اعمال معمولا به سمت آنچه کسب و کار انجام می دهد، پیش می‌روند:

1- ارسال پیام ها به مشتریان و انتظارات از طریق پست مستقیم، پست الکترونیک، بازاریابی از راه دور و غیره.
2- با داده‌کاوی، این امکان بوجود می آید که پیام‌های مختلف به افراد مختلفی فرستاده شود.
3- اولویت بندی سرویس های مشتریان
4- تعدیل کردن سطوح موجودی
5- وغیره

لازم است که نتایج داده کاوی روال های کسب و کار را تغذیه کند تا مشتریان لمس شوند و روی روابط با مشتریان تاثیر گذاشته شود.

4- اندازه گیری نتایج :

اهمیت اندازه گیری نتایج، در حال حاضر پررنگ خواهد شد. علی‌رغم اهمیت آن، این مرحله ای است در چرخه تعالی که اغلب چشم‌پوشی می شود. حتی با اینکه بر ارزش اندازه‌گیری و بهبود پیوسته، بسیار تاکید شده است، معمولا به نسبت استحقاق آن، توجه کمتری به آن می شود. چقدر موارد کسب و کاری وجود دارد که پیاده سازی شده اند و هیچ کسی برنگشته تا ببیند آیا آنچه پیاده‌سازی شده واقعا با طرح سازگار است یا خیر؟ افراد تلاش هایشان را با مقایسه و یادگیری کنترل می‌کنند، با طرح سوالاتی درباره اینکه چرا طرح ها با آنچه واقعا اتفاق افتاده سازگار است یا سازگار نیست. آنچه برای افراد صدق می‌کند برای سازمان ها نیز کاربرد دارد.

زمان شروع تفکر درباره اندازه‌گیری در شروع زمان شناسایی مسئله کسب و کار است. چطور می‌توان نتایج را اندازه‌گیری کرد؟ یک معیار مناسب دیگر، سنجش افزایش فروش در مغازه‌ها و یا مناطق مشخص است. این افزایشها را می‌توان به تلاشهای خاص بازاریابی مرتبط نمود. ممکن است چنین سنجش‌هایی از آنجا که نیازمند اطلاعات جزئی فروش است مشکل باشد. با این وجود اگر هدف افزایش فروش است روشی برای بررسی مستقیم آن لازم است.

ایده خوبی‌است که به هر تلاش داده کاوی به عنوان یک مورد کوچک تجاری نگاه کنیم.

مقایسه انتظارات با نتایج عملی این امکان را ایجاد میکند که فرصت‌های آتی را تشخیص دهیم تا از آنها در دوره های بعدی چرخه تعالی بهره ببریم. هر اقدام داده‌کاوی چه موفقیت‌آمیز باشد یا نه، حاوی نکاتی است که می‌تواند برای تلاش‌های آتی بکار رود.

سؤال این است که چه چیزی را بررسی نماییم و چگونه به آن بپردازیم، تا بهترین ورودی‌ها را برای استفاده های آتی داشته باشیم.

۳-۲ فرایند داده کاوی:

داده کاوی نیازمند شناسایی یک مسئله همراه با مجموعه ای از داده هاست که می تواند منجر به درک بهترومدل های کامپیوتری به منظورفراهم اوردن تحلیل اماری یا مقاصد دیگرگردد. این کار با ابزارهای تجسم فکری که داده را نمایش می دهد یا تحلیل اماری کاملاً بنیادی مثل تحلیل همبستگی ممکن است پشتیبانی شود .

ابزارهای داده کاوی نیاز دارند که متنوع ، درجه بندی شده ، قادر به واکنش های پیشگویی دقیق بین عملیات و نتایج و اجرای خودکارباشند . تنوع به توانایی ابزاری اشاره دارد که انواع وسیعی ازمدل ها را به کار گیرد . ابزارهای مقیاس اشاره به این موضوع می کنند که اگراین ابزارها روی یک مجموعه داده کوچک کارکند باید روی مجموعه داده های بزرگتر نیز کار کند . کنترل خودکار سودمند است اما کاربردش وابسته به اموری است .

بعضی کارکردهای تحلیلی اغلب خودکار هستند اما تنظیمات بشری مقدم بر اجرای رویه ها مورد نیاز هستند . در حقیقت قضاوت تحلیل گر انتقادی به اجرای موفقیت امیزداده کاوی است . انتخاب مناسب داده ها که شامل تحقیقات می باشد حیاتی است . تغییر شکل داده نیز اغلب مورد نیاز است .

متغیرهای بسیاری بازده زیادی تولید می کنند . در حالی که تعداد کمی می توانند بر روابط کلیدی در داده ها مسلط شوند . درک بنیادی مفاهیم اماری برای داده کاوی موفق ضروری است .

داده کاوی به سرعت با فواید زیادی برای تجارت بسط داده می شود . دو ناحیه کاربرد با بیشترین سود ، استفاده از تقسیم مشتری از طریق سازمان های بازاریابی برای تشخیص احتمالات بیشتر حاشیه ای واکنش به شکل های متفاوت رسانه بازاریابی ، و بانک ها با استفاده از داده کاوی برای پیش گویی دقیق تر احتمال این که افراد به پیشنهاداتی که سرویس های مختلف ارائه می کنند واکنش می دهند ، می باشد .

بسیاری از شرکت ها از این تکنولوژی برای شناسایی مشتریان استفاده می کنند تا اینکه بتوانند خودشان سرویسی را که نیاز است فراهم اورند تا ان مشتریان را از دست ندهند .

متدلوژی داده‌کاوی و بهترین تمرین‌های آن:

در بخش قبل ما چرخه تعالی داده‌کاوی و مراحل آنرا بررسی کردیم. حالا زمان آن رسیده که به داده‌کاوی بعنوان یک روال تکنیکی نگاهی بیندازیم. رئوس مطالب در سطح بالا، همان باقی می‌ماند ولی تاکیدها منتقل می‌شود. به جای شناسایی مسئله کسب و کار، حالا می خواهیم توجه‌مان را از مسئله کسب و کار به مسئله داده‌کاوی منتقل کنیم.

بهترین راه برای دوری از شکست چرخه داده‌کاوی، درک راه‌های عدم موفقیت آن و اخذ تدابیر پیشگیرانه است. در طول سالها، مولفان با راه‌های زیادی برای پروژه‌های داده‌کاوی مواجه شده‌اند که به اشتباه رفته اند. در پاسخ، ما مجموعه ای از عادات کارآمد را توسعه داده ایم- چیزهایی که ما باید برای هموار کردن مسیرمان از جمله ابتدایی مسئله کسب و کار تا یک مدل پایدار که نتایج قابل اندازه‌گیری و قابل اعمالی را تولید می‌کند، انجام دهیم. داده‌کاوی یک روال تکراری طبیعی است. لازم است برخی مراحل را چندین بار تکرار کرد، ولی هیچیک را نمی‌توان کاملا نادیده گرفت.

با افزایش پیچیدگی راه حل داده‌کاوی، نیاز به وجود یک مشی سفت و سخت بیشتر است. پس از اینکه نیاز به وجود متدلوژی با راه‌های مختلفی توضیح داده شد، با استفاده از پرسش های خاص منظوره نشان داده خواهد شد که در صورت نبود یکی از مراحل آن، تلاش های داده‌کاوی با شکست مواجه خواهند شد. در پایان، 4 مرحله از چرخه تعالی داده‌کاوی به 11 مرحله متدلوژی داده‌کاوی تبدیل می‌شود.

چرا یک متدلوژی داشته باشیم؟

داده‌کاوی راهی است برای یادگیری از گذشته تا بتوان تصمیمات بهتری در آینده گرفت. داشتن یک متدلوژی برای اجتناب از دو نتیجه نامطلوب روال یادگیری است:

1- یادگیری چیزهایی که درست نیستند
2- یادگیری چیزهایی که درستند ولی کارآمد نیستند

یادگیری چیزهایی که درست نیستند بسیار خطرناک‌تر از یادگیری چیزهایی است که کارآمد نیستند، چون بسیار مهم است که تصمیمات سازمان براساس اطلاعات درست گرفته شود. یافته‌های داده‌کاوی به نظر می‌رسد که قابل اعتماد باشند چون بر اساس داده‌های واقعی در یک حالت ظاهرا علمی هستند. این نمود قابل اعتماد می‌تواند فریبنده باشد. ممکن است داده نادرست باشد یا مرتبط با سوال نباشد. انتقال داده‌ها مانند خلاصه‌سازی‌ها، ممکن است مشکل‌دار باشند یا اطلاعات مهمی را پنهان کنند. در زیر درباره برخی از مسائل بسیار رایجی که منجر به استنتاجات اشتباه می‌شوند، بحث می‌کنیم.

الگوهایی که ممکن است هیچ قانون اصولی را ارائه نکنند :

ارقام اغلب دروغ نمی‌گویند، ولی دروغ ها می‌توانند رقم شوند. وقتی می‌خواهیم الگوها را در داده‌ها بیابیم، ارقام واقعا نمی‌خواهند با هدف پیشنهاد چیزهای نادرست، بما دروغ بگویند. راه‌های زیادی وجود دارد که الگوهایی ساخته شود که هر مجموعه داده تصادفی، اگر به اندازه کافی آزمایش شود، یکی را آشکار خواهد کرد. ما انسانها، به الگوها در زندگی‌مان بسیار وابسته‌ایم و دوست داریم آنها را همه جا ببینیم حتی اگر وجود نداشته باشند. وقتی به آسمان شب نگاه می‌کنیم و یک مجموعه تصادفی ستارگان را می‌بینیم، آنها را به شکل دب اکبر، دب اصغر و غیره می بینیم. گاهی حتی الگوهای مربوط به نجوم را نشانه‌هایی می‌بینیم که از آینده خبر می‌دهند.

احتمالا دلیل اینکه بشر این قبیل وابستگی‌ها را برای الگوها استنتاج می‌کند این است که این الگوها اغلب برخی حقایق را درباره کار جهان، بازتاب می‌دهند.

چالش پیش روی کاونده داده این است که کشف کند که کدام الگوها پیشگویانه هستند و کدام نیستند. به الگوهای زیر دقت کنید، همه آنها در مقالات بعنوان اینکه ارزش پیش بینی دارند، ذکر شده‌اند:

1- حزبی که در ریاست جمهوری را برنده نشده، در انتخابات off-year کرسی‌های کنگره را برنده می‌شود.
2- وقتی که لیگ امریکایی سری‌های جهانی را برنده می‌شود، جمهوری‌خواهان کاخ سفید را می برند.
3- در مناظرات ریاست جمهوری آمریکا، کسی که قدبلندتر است، معمولا برنده است.

الگوی اول، به نظر می‌رسد که با عبارات صرف سیاسی قابل توضیح باشد. چون در اینجا یک توضیح متضمن آن وجود دارد، این الگو به نظر می‌رسد به آینده نظر دارد بنابراین ارزش پیشگویانه خواهد داشت. دو عبارت بعدی، یکی که حاوی حوادث ورزشی است، به نظر می‌رسد که به وضوح ارزش پیشگویی نداشته باشد. هیچ اهمیتی وجود ندارد که چند بار در گذشته فاتحان جمهوری خواهان و لیگ آمریکایی، مشترک بوده اند و این چیزی است که تحقیقی برای آن انجام نشده است. در مورد قد کاندیدها چطور؟ از سال 1945 تا کنون، همیشه کاندیداهای قد بلندتر، برنده انتخابات بوده‌اند. اما به نظر می‌رسد که قد هیچ ربطی به ریاست جمهوری نداشته باشد. از سوی دیگر، قد، ارتباط مثبتی روی درآمد و سایر موفقیت‌های اجتماعی دارد، و انتخاب‌کنندگان، بطور عمدی یا غیرعمدی، کاندیداهای بلندتر را ترجیح می‌دهند.

چیدمان مدل ممکن است بازتاب دهنده جمعیت وابسته نباشد :

چینش مدل، مجموعه ای از داده‌های تاریخی است که برای توسعه مدل داده‌کاوی بکار می‌رود. برای اینکه تفاسیر برگرفته از چینش مدل، معتبر باشد، چیدمان مدل باید منعکس کننده جمعیتی باشد که مدل قصد تشریح، طبقه بندی یا رتبه‌بندی آنرا داشته است. یک مثال که جمعیت والد به خوبی منعکس نشده است، جانبداری است. استفاده از یک نمونه جانبدارانه، بعنوان یک چیدمان

مدل، یک دستورالعمل برای موارد آموزشی‌ای است که درست نیستند. اجتناب از آن نیز سخت است. فرض کنید:

1- مشتریانی که مطابق با انتظارات نیستند.
2- ممیزی پاسخ دهندگان مانند آنها که پاسخ نداده‌اند نیست.
3- افرادی که Email می خوانند مانند آنها که نمی‌خوانند نیستند.
4- افرادی که در یک سایت وب ثبت‌نام می‌کنند، مانند آنها که موفق به ثبت‌نام نمی‌شوند، نیستند.
5- رکوردهایی که هیچ مقدار گم شده ای ندارند، جمعیت متفاوتی را به نسبت رکوردهایی که مقادیر گم شده دارند، منعکس می‌کنند.

در انتخاب و نمونه‌گیری از داده‌ها برای مدل، بسیار دقیق باشید؛ زیرا در موفقیت داده‌کاوی بسیار موثر خواهد بود.

یادگیری چیزهایی که درست ولی بلااستفاده‌اند :

یادگیری چیزهایی که بلااستفاده اند، اگرچه به خطرناکی یادگیری چیزهایی که نادرستند نیست، ولی بسیار رایج است.

1- یادگیری چیزهایی که در حال حاضر می دانیم

داده‌کاوی باید اطلاعات جدیدی تولید کند. بسیاری از الگوهای قوی در داده، چیزهایی را که می‌دانیم، ارائه می‌کنند. مثلا افراد بالای سن بازنشستگی، تمایلی به پاسخ دادن به طرح های ذخیره بازنشستگی، ندارند. افرادی که در جایی زندگی می‌کنند که برج مخابرات ندارند، تمایلی به خرید تلفن ندارند.

قویترین الگوها، اغلب قانون‌های کسب و کار را بازتاب می‌دهند. اگر داده‌کاوی کشف کند که افرادی که امکان قطع تلفن‌های نامشخص را دارند، caller ID نیز دارند، این محتمل است چون که این امکان با بسته‌ای فروخته می‌شود که شامل caller ID نیز هست. ما بسیاری از این الگوهای کشف شده را در داده‌کاوی می بینیم. این الگوها نه تنها مطلوب نیستند، که ممکن است قدرت آنها، الگوهای بدیهی دیگری را محو کند.

یادگیری چیزهایی که درحال حاضر می دانیم، یک هدف مفید دارد اینکه ثابت می‌کند که در سطح تکنیکی، تلاش داده‌کاوی کار می‌کند و داده صحیح است. این می‌تواند کاملا تسلی بخش باشد. اگر داده و داده‌کاوی اعمال شده بر آن به اندازه کافی قدرتمند باشد که چیزهای شناخته شده را بدرستی کشف می‌کند، پس این اطمینان را می‌دهد که سایر کشفیاتش نیز درست باشد.

2- یادگیری چیزهایی که نمی‌توان آنها را مورد استفاده قرار داد

این زمانی اتفاق می افتد که داده‌کاوی روابط پوشش داده نشده‌ای را که هم درستند و هم قبلا شناخته نشده‌اند را کشف می‌کند ولی مشکل است که از آنها استفاده کرد. مثلا سابقه اعتباری یک مشتری ممکن است یک دعوی بیمه را در آینده پیش بینی کند، اما تنظیم کننده، اتخاذ تصمیم را بر مبنای آن منع کند.

داده‌کاوی ممکن است خروجی های دیگری را که خارج از کنترل شرکت هستند را پیش بینی کند. ممکن است یک محصول برای یک آب و هوا مناسب تر از آب و هوای دیگری باشد، ولی سخت است که آب و هوا را کنترل کرد.

نکته: بعضی اوقات ممکن است یک تصور اشتباه باعث شود که اطلاعات جدید را بی‌استفاده بدانیم. پیش‌بینی‌های رویگردانی مشتریان ممکن است دیگر برای بکارگیری حفظ مشتریان کنونی بسیار دیر باشد، اما می‌تواند ما را برای یافتن راههایی برای تغییر کانال های ارتباطی مان با مشتریان آینده، ترغیب کند.

2-3-1مدل‌ها، پروفایل‌سازی، و پیش‌بینی :

تکنیک‌های داده‌کاوی شرح داده شده در اینجا، همگی برای یادگیری چیزهای جدید با ساخت مدل‌ها بر اساس داده‌ها، طراحی شده‌اند.

سراسر داده‌کاوی درباره ساختن مدل‌هاست. همانطور که در شکل 4 نشان داده شده است، مدل‌ها مجموعه‌ای ورودی می‌گیرند و یک خروجی تولید می‌کنند. داده‌ای که برای ساختن مدل بکار می‌رود مجموعه مدل نامیده می‌شود. و هنگامی که مدل‌ها روی داده‌های جدید اعمال می‌شوند، مجموعه رتبه نامیده می‌شوند. مجموعه مدل سه جزء دارد:

1- مجموعه آموزشی که برای ساختن یک مجموعه مدل بکار می‌رود.
2- مجموعه اعتبارسنجی که برای انتخاب بهترین مدل از میان اینها بکار می‌رود.
3- مجموعه آزمون که برای تعیین چگونگی عملکرد مدل بر روی داده‌های دیده نشده، بکار می‌رود.

تکنیک‌های داده‌کاوی را می‌توان برای ساختن سه نوع مدل برای سه نوع وظیفه، استفاده کرد: پروفایل‌سازی توصیفی، پروفایل‌سازی جهتدار، و پیش‌بینی.

مدل‌های توصیفی، تشریح می‌کنند که چه چیزی در داده است. خروجی آنها، نمودارها یا اعداد یا گرافیک‌هاست. از سوی دیگر، پروفایل سازی جهتدار و پیش‌بینی، از ساخت مدل، هدفی را دنبال می‌کنند. تفاوت این دو مدل در چارچوب‌های زمانی است که در شکل شان داده شده است. در مدل پروفایل‌سازی، هدف در همان چارچوب زمانی ورودی قرار دارد در حالیکه در مدل پیش‌گویی، هدف در چارچوب زمانی بعدی است.

2-3-2 پروفایل‌سازی :

پروفایل‌سازی یک دستاورد آشنا برای بسیاری از مسائل است. نیازی به هیچ تحلیل داده سطح بالا و پیچیده ای ندارد. پروفایل‌ها، اغلب بر اساس متغیرهای سرشماری، مثل جنسیت، سن و مکان جغرافیایی هستند. از زمانی که تبلیغات بر طبق این متغیرها فروخته می‌شوند، پروفایل‌های سرشماری می‌توانند مسقیما در استراتژی های رسانه قرار گیرند. پروفایل های ساده ای برای حق بیمه شرکتهای بیمه بکار می روند. یک مرد 17 ساله، بنسبت یک زن 60 ساله، حق بیمه بیشتری برای ماشینش باید پرداخت کند. همچنین کاربردهایی که برای پرداخت حق بیمه عمر هستند، سوالاتی درباره سن، جنسیت و سیگاری بودن می پرسند.

علی‌رغم این قدرت، پروفایلینگ محدودیت‌های جدی‌ای دارد. اول اینکه قادر به تشخیص انگیزه و نتیجه نیست. مادامی که پروفایل‌سازی مبتنی بر متغیرهای سرشماری آشنا باشد، این نمی‌تواند قابل توجه باشد.

بـا داده‌هـای رفتـاری، مسـیر علیـت همیشـه روشـن نیسـت. یـک جفـت مثـال واقعـی از پـروژه‌هـای داده‌کاوی واقعی را فرض کنید:

1- افـرادی کـه بـا سـپرده‌هـای اعتبـاری شـان خریـد کـرده انـد، در حسـاب‌هـای ذخیـره‌شـان، مقدار کم یا هیچ پولی ندارند.
2- مشتریانی که از پسـت صـوتی اسـتفاده مـی‌کننـد، تمـاس‌هـای کوتـاه زیـادی بـا شـماره‌شـان می‌گیرند.

نگه نداشتن پول در حسـاب‌هـای ذخیـره، رفتـار رایـج دارنـدگان سـپرده‌هـای اعتبـاری اسـت. مثـل مـرد بودن که خصوصیت رایج نوشـندگان نوشـابه اسـت. شـرکت هـای نوشـابه، مشـتریان مـرد را بـرای بـازار محصولشان جسـتجو مـی‌کننـد، همینطـور بانـک هـا، بـرای فـروش سـپرده‌هـای اعتبـاری‌شـان، کسـانی را حساب ذخیره خالی دارنـد، جسـتجو مـی‌کننـد؟ شـاید نـه! احتمـالا دارنـدگان سـپرده‌هـای اعتبـاری هـیچ پـولی در حسـاب هـای ذخیـره خـود ندارنـد چـون پولشـان را بـرای خریـد از اعتبارشـان اسـتفاده مـی‌کننـد. یـک دلیـل رایـج بـرای نداشـتن پـول در حسـاب هـای ذخیـره، نداشـتن هـیچ پـولی اسـت و کسانی که پولی ندارند برای خریدن کارت اعتباری اصلا مطلوب نیستند.

8-1-2 پیش بینی :

پروفایـل سـازی، داده‌هـای گذشـته را بـرای تشـریح آنچـه کـه در گذشـته رخ داده اسـت بکـار مـی‌بـرد. پیش‌بینی یک گـام بـه جلوسـت. و داده‌هـای گذشـته را بـرای پـیش‌بینـی آنچـه کـه در آینـده رخ خواهـد داد، بکار می‌برد. این استفاده از داده، قدرت بیشتری دارد.

ساختن مدل پیشـگویانه، نیـاز بـه جـدایی زمـانی بـین ورودی هـای مـدل یـا پیشـگو، و خروجـی مـدل، چیزی کـه پـیش بینـی شـده، دارد. اگـر ایـن جـدایی حاصـل نشـود،مدل کـار نخواهـد کـرد. ایـن مثـالی است از اینکه چرا مهم است که از متدلوژی داده‌کاوی پیروی کنیم.

٤-٢ متدلوژی :

متدلوژی داده‌کاوی 11 مرحله دارد:

1- تبدیل مسئله کسب و کار به مسئله داده‌کاوی

2- انتخاب داده مناسب
3- رسیدن به شناخت داده
4- ایجاد یک مجموعه مدل
5- تثبیت مسئله با داده
6- تبدیل داده برای آوردن اطلاعات به سطح
7- ساختن مدل ها
8- ارزیابی مدل ها
9- استقرار مدلها
10- ارزیابی نتایج
11- شروع دوباره

مرحله 1: تبدیل مسئله کسب و کار به مسئله داده‌کاوی :

اهداف داده‌کاوی برای یک پروژه خاص نباید با عبارت کلی نوشته شده باشد، مثلا:

1- کسب درکی از رفتار مشتری
2- کشف داده‌های معنی‌دار از داده‌ها
3- یادگیری چیزهای جالب

تمام اینها اهداف شایسته‌ای هستند ولی حتی هنگامی که آنها بدست آیند، اندازه گیری آنها مشکل است. پروژه‌هایی که اندازه‌گیری آنها سخت است، ارزش‌گذاری آنها سخت است. تا آنجا که ممکن است، باید اهداف کلی را به اهداف خاص دیگری شکسته شوند که نظارت بر روال دسترسی به آنها آسانتر شود. بدست آوردن درکی از رفتار مشتری، می‌تواند به اهداف زیر شکسته شود:

1- شناسایی مشتریانی که دوست ندارند اشتراکشان را تجدید کنند.
2- طراحی یک طرح مکالمه‌ای که رویگردانی مشتریان خانگی را کاهش دهد
3- رتبه بندی تمام مشتریانی که گرایش به اسکی دارند.

این اهداف واقعی، فقط برای نظارت آسانتر نیستند بلکه تبدیل آنها به مسائل داده‌کاوی نیز آسانتر است.

برای تبدیل یک مسئله کسب و کار به یک مسئله داده‌کاوی، باید آنرا بصورت یکی از وظایف داده‌کاوی شرح داده شده دربخش قبل، دوباره فرمول سازی کرد

از نتایج چگونه استفاده می‌شود ؟

این یکی از سوالات بسیار مهمی است که در این مرحله باید از خود بپرسیم. بطور شگفت انگیزی، پاسخ در ابتدا این است: مطمئن نیستم! اما پاسخ به این سوال بسیار مهم است چون کاربردهای مختلف، راه‌حل‌های مختلفی را نیز پیشنهاد می‌کنند.

نتایج چگونه تحویل داده می‌شوند ؟

پروژه‌های داده‌کاوی مختلف، ممکن است چندین نوع نتایج قابل تحویل داشته باشند. مثلا وقتی هدف اوله پروژه، بدست آوردن درکی از مشتری باشد، نتیجه قابل تحویل اغلب یک گزارش، چارت یا نمودار یا گراف است. شکل قابل تحویل می‌تواند نتایج داده‌کاوی را تحت تاثیر قرار دهد.

نقش کاربران کسب و کار و تکنولوژی اطلاعات :

تنها راه پاسخ دادن مطلوب به سوالات مطرح شده در بالا، درگیر کردن صاحبان کسب و کار در کشف چگونگی استفاده از نتایج داده‌کاوی، و درگیر کردن کارکنان فن‌آوری اطلاعات در کشف نحوه تحویل این نتایج است.

مرحله 2: انتخاب داده مناسب :

داده‌کاوی به داده نیاز دارد. در بهترین حالت ممکن، داده مورد نیاز در یک انبار داده مجتمع، پالایش شده، در دسترس، با سابقه درست، و بطور متناوب در حال اصلاح، قرار دارد. اما در واقعیت تمام این موارد ممکن نیست. منابع داده مفید و دردسترس، از مسئله‌ای به مسئله دیگر، و از صنعتی به صنعت دیگر، متنوع هستند.

چه چیزی در دسترس است ؟

اولین جایی که باید بدنبال داده گشت، یک انبار داده مجتمع است. داده‌ها در انبار، پالایش، اعتبارسنجی و از چندین منبع باهم گردآوری شده‌اند. یک مدل داده تکی، این اطمینان را می‌دهد که فیلدها بطور مشابهی نامگذاری شده‌اند، معنای یکسانی، وانواع داده سازگاری را در پایگاه داده دارند. یک انبار داده یکی شده، یک مخزن سابقه ای است؛ داده‌های جدید به آن اضافه می‌شوند ولی داده‌های قبلی تغییر داده نمی‌شوند. از آنجایی که برای پشتیبانی تصمیم گیری طراحی شده‌اند، انبار داده، داده‌های با جزئیاتی را که برای داده‌کاوی در سطح درستی پذیرفته شود، تهیه می‌کند. تنها مسئله این است که در بسیاری از سازمان ها، واقعا چنین انبار داده ای وجود ندارد یا اینکه یک یا چند انبار داده وجود دارد ولی در سطح قابل قبولی نیستند. در شروع این مورد، داده‌کاو باید داده‌ها را از میان پایگاه داده‌های مختلف اداری و از دل سیستمهای عملیاتی مختلف بیرون بکشد. ممکن است سیاست‌های معنی‌دار و تلاش‌های برنامه نویسی‌ای برای گرفتن داده بشکل کارآمد برای کشف دانش از چنین سیستم هایی نیاز شود. در برخی موارد، روال های عملیاتی برای تهیه داده، تغییر خواهند کرد.

چه مقدار داده کافی است ؟

متاسفانه پاسخ ساده ای برای این سوال وجود ندارد. پاسخ به الگوریتم های خاصی که بکار گرفته می‌شود و پیچیدگی داده، بستگی دارد. در مواردی که داده کم است، داده‌کاوی نه تنها کم اثر است بلکه بلااستفاده نیز هست. داده‌کاوی بیشترین فایده را زمانی دارد که حجم خالصی از داده، الگوهایی را پنهان کند که قابل جستجو در پایگاه داده‌های کوچک هستند.

در داده‌کاوی، داده‌ها هر چه بیشتر، بهتر است ولی باید به دو نکته توجه کرد. اولین نکته، رابطه بین اندازه مجموعه مدل و تراکم آن است. منظور از تراکم پخش اثرات مطلوب‌هاست. اغلب متغیر هدف، چیزی را که نسبتا نادر است، ارائه می‌کند. برای دارندگان کارت های اعتباری، ارتکاب تقلب نادر است. اینکه مشترکان یک روزنامه، اشتراکشان را لغو کنند، کم پیش می‌آید. برای مجموعه مدل، مطلوب این است که با اعداد برابر با هر یک از خروجی های روال ساخت

مـدل، متـوازن باشـد. یـک مجموعـه کـوچکتر متـوازن بـه مجموعـه بزرگتـری کـه مـوارد نـادر را داشـته باشد، ترجیح داده می‌شود.

دومین نکته، این اسـت کـه وقتـی مجموعـه مـدل بـرای سـاختن مـدل هـای خـوب و پایـدار، بـه انـدازه کـافی بـزرگ اسـت، بزرگتـر سـاختن آن خنثـی کننـده تولیـد اسـت، چـون هـر چیـزی بـرای اجـرا روی مجموعه داده، زمان بیشتری مـی بـرد. از آنجـا کـه روال داده‌کـاوی یـک فراینـد تکـراری اسـت، اگـر هـر اجرا از یک روتین مدل‌سازی بـه جـای چنـد دقیقـه، سـاعت هـا وقـت بگیـرد، زمـان صـرف شـده بـرای رسیدن به نتایج ممکن است طولانی شود.

یک تست ساده بـرای اینکـه بـدانیم مجموعـه داده‌هـا بـرای شـروع بـه انـدازه کـافی مناسـب اسـت ایـن است که انـدازه نمونـه را دوبرابـر کنـیم و بهبـود درسـتی مـدل را انـدازه گیـری کنـیم. اگـر مـدل سـاخته شده از داده‌های بزرگتـر، بطـرز معنـی‌داری بهتـر از مـدل کـوچکتر بـود، بـدان معنـی اسـت کـه مجموعـه نمونه ما به اندازه کافی بـزرگ نیسـت. امـا اگـر بهبـود چشـمگیری در مـدل حاصـل نشـد، احتمـالا مـدل اصلی مناسب خواهد بود.

چقدر سابقه لازم است ؟

داده‌کاوی، داده‌های گذشته را برای پیش بینی آینـده بکـار مـی بـرد. ولـی چـه مقـدار از سـابقه را بایـد بـا داده‌ها آورد؟ این یـک سـوال سـاده دیگـری اسـت کـه پاسـخ سـاده ای نـدارد. اولـین چیـز بـرای توجـه فصلی بـودن اسـت. برخـی کسـب و کارهـا درجـه ای از فصـلی بـودن را دارنـد. مـثلا سـفرهای فراغـت در تابستان اتفاق می‌افتد. فروش‌هـا بیشـتر در یـک چهـارم پایـانی سـال انجـام مـی‌شـود. بایـد بـه انـدازه کافی داده قبلی برای رویدادهای دوره‌ای از این دست، گرفته شود.

از سوی دیگـر، داده‌هـای گذشـته‌هـای دور بـدلیل تغییـر شـرایط بـازار، نمـی‌تواننـد خیلـی مفیـد باشـند. برای بیشتر کاربردهای متمرکز برمشتری، سوابق دو تا سه سال، مناسب است.

چند تا متغیر ؟

داده‌کاوهـای بـی‌تجربـه، بعضـی اوقـات در دور ریخـتن متغیرهـایی کـه بـه نظرشـان بـدردنخور هسـتند و نگه داشتن تعـداد کمـی از متغیرهـایی کـه بـه نظرشـان مهـم مـی‌رسـند، بسـیار شـتاب دارنـد. روش‌هـای

داده‌کاوی این اجازه را می‌دهند که داده خودش مشخص کند که چه چیز مهم و چیز بی‌اهمیت است. اغلب، متغیرهایی که قبلا از آنها چشم پوشی کرده‌ایم و آنها را کنار گذاشته‌ایم، در صورت ترکیب با متغیرهای دیگر، ارزش پیشگویانه پیدا می‌کنند. این درست است که مدل نهایی فقط براساس تعداد کمی متغیر بنا نهاده می‌شود. ولی این متغیرهای اندگ گاهی از ترکیب چندین متغیر دیگر بدست می‌آیند و ممکن است که در ابتدا قابل مشاهده نباشند که یکی از آنها در نهایت مهم خواهد بود.

داده باید حاوی چه چیزی باشد ؟

داده در کمترین حالت باید حاوی تمام مثال های ممکن از نتایج مطلوب باشد. در داده‌کاوی مستقیم، که هدف پیش بینی متغیر هدف است، داشتن یک مجموعه مدل دربردارنده داده‌های دسته‌بندی شده، بسیار حساس است. برای تمایز بین افرادی که دوست دارند وام دریافت کنند و افرادی که دوست ندارند، از هر دو دسته مثال‌های متفاوت زیادی لازم است تا مدلی ساخته شود که یکی را از دیگری تشخیص دهد. وقتی یک درخواست جدید می‌رسد، درخواست او با مشتریان گذشته مقایسه می‌شود، یا بطور مستقیم در استدلال مبتنی بر حافظه، یا بطور غیرمستقیم از طریق قوانین یا شبکه‌های عصبی برگرفته از داده‌های سابقه‌ای. اگر درخواست جدید "شبیه به نظر رسید" با درخواست هایی که در گذشته قرارداد شده‌اند، رد خواهد شد.

مرحله 3: پیش به سوی شناخت داده :

داده‌کاوان خوب، بر روی درک مستقیم خیلی تکیه می‌کنند. تنها راه برای توسعه درک یک مجموعه داده ناآشنا، غوطه‌ور شدن در آن است. از این راه شما بسیاری از مسائل کیفیت داده را کشف خواهید کرد و سوالات بسیاری برایتان مطرح می‌شود.

آزمودن توزیع‌ها :

به هر چیزی که شما را متحیر می‌کند، توجه کنید؛ اگر یک متغیر کد ناحیه داریم، آیا "کالیفرنیا"، بیشترین طول را دارد؟ اگر متغیر جنسیت داریم، آیا تعداد زنان و مردان یکی است؟ به دامنه هر

متغیری دقت کنید. ابزارهای بصری‌سازی، می‌تواند در راه کنکاش اولیه پایگاه‌داده، کمک‌کننده باشد.

مقایسه مقادیر با توصیف‌ها :

به مقادیر هر یک متغیرها دقت کنید و مقایسه کنید که آیا با توصیف آن متغیر در مستندات، مطابقت دارد؟ این تمرین مشخص می‌کند که آیا مستندات درست و کامل هستند؟

اعتبارسنجی فرضیات :

با استفاده از ابزارهای بصری‌سازی، مثل گراف‌های خطی، نگاشت‌ها و نقشه‌های پراکنده، می‌توان فرضیاتی را درباره داده‌ها، بررسی و اعتبارسنجی کرد. به رابطه متغیر مقصد و سایر متغیرها دقت کنید تا چیزهایی از این قبیل را ببینید، مثل رابطه بین درآمد و جنسیت.

پرسیدن سوالات بسیار:

یک خروجی مهم از روال جستجوی داده‌ها، فهرستی از سوالاتی است از افرادی که داده‌ها را تهیه می‌کنند. اغلب این سوالات، بررسی‌های بیشتری را می‌طلبد، چون کاربران کمی، به همان دقت داده‌کاو، داده را بررسی می‌کنند.

مرحله 4: ساختن یک مجموعه مدل :

مجموعه مدل، شامل تمام داده‌هایی است که در فرایند مدلسازی بکار می‌روند. یکسری از این داده‌ها، برای یافتن الگوها، برخی دیگر برای اعتبارسنجی مدل، و برخی دیگر برای ارزیابی کارایی مدل، بکار می‌روند. ساختن یک مجموعه مدل، نیاز به اسمبل کردن داده از منابع چندگانه و سپس آماده کردن آنها برای تحلیل، دارد.

گردآوری امضاهای مشتری :

مجموعه مدل، یک جدول یا مجموعه‌ای از جداول با یک ردیف برای هر آیتم و فیلدهایی برای هر آنچه درباره آن آیتم می‌دانیم، است. وقتی داده مشتریان را توصیف می‌کند، ردیف‌های

مجموعه مدل اغلب "امضاهای مشتری" نامیده می‌شوند. گردآوری این اطلاعات اغلب نیاز به پرسش‌های پیچیده از چندین جدول و سپس یکپارچه‌سازی آنها با داده‌های دیگر منابع، دارد.

ساختن یک نمونه متعادل :

اغلب وظیفه داه‌کاوی شامل یادگیری تمایز بین گروه‌های مختلف است: مثل پاسخگوها و پاسخ‌ندهندگان، خوب‌ها و بدها، و غیره. اگر تعداد اعضای این گروه‌ها متناسب باشد، الگوریتم‌های داده‌کاوی بهتر عمل می‌کنند. بنابراین قبل از مدلسازی، مجموعه داده باید با نمونه‌سازی از گروه‌های مختلف با نرخ‌های متفاوت و افزودن یک فاکتور وزن دهی، متوازن شود.

شامل چارچوب‌های زمانی چندگانه باشد :

هدف اصلی یک متدلوژی، ساختن یک مدل پایدار است. یکی از معانی آن این است که مدل‌ها برای هر زمانی از سال کار کنند و برای آینده مناسب باشند. این اتفاق زمانی می افتد که در مجموعه مدل، تمام داده‌ها از یک دوره خاص زمانی از سال، نیامده باشند. ساختن مدل براساس یک دوره زمانی خاص، این خطر را دارد که الگوهایی که عموما درست نیستند را آموزش دهد. یک مثال جالب برای این مسئله این است که یکبار یک قانون وابستگی را برای داده‌های یک هفته سوپرمارکتی بدست آوردند که تمام قانون‌ها به تخم مرغ، ختم می‌شد. یعنی پیش‌بینی سبد خرید، بیان می‌کرد که هر کسی چیزی می‌خرد، تخم‌مرغ هم می‌خرد! البه این نباید زیاد تعجب برانگیز باشد اگر بدانیم که مجموعه مدل ما مربوط به هفته قبل از عید پاک بوده است!

ساختن یک مجموعه مدل برای پیش‌بینی :

اگرچه ممکن است که مجموعه مدل شامل فریم‌های مختلف زمانی باشد، اما هر امضای مشتری، یک فاصله زمانی بین متغیرهای پیش‌بینی کننده و متغیرهای هدف، خواهد داشت. همیشه می‌توان زمان را به سه دوره تقسیم کرد: گذشته، حال، آینده. وقتی یک پیش‌بینی انجام می‌شود، مدل، داده‌های گذشته را برای پیش‌بینی آینده بکار می‌گیرد. تمام این سه دوره زمانی باید در مجموعه مدل نشان داده شود. البه تمام داده‌ها در پایگاه‌دادها از گذشته می‌آیند؛ بنابراین دوره‌های زمانی

در مجموعه مدل، بطور دقیق‌تر شامل گذشته دور، گذشته نچندان دور، و گذشته اخیر است. مدلهای پیش‌بینی برای یافتن الگوهایی در گذشته دور ساخته شده‌اند که خروجی‌هایشان، گذشته اخیر را توصیف می‌کند. وقتی مدل مستقر شود، می‌تواند داده‌های گذشته اخیر را برای پیش‌بینی آینده بکار برد.

اگر یک مدل، داده‌های ماه June(گذشته نه چندان دور) را برای پیش‌بینی July(گذشته اخیر) بکار برد، نمی‌تواند تا زمانی که داده‌های آگوست در دسترس باشند، برای پیش‌یبنی سپتامبر استفاده شود. ولی داده‌های آگوست کی در دسترس خواهد بود؟ قطعا در خود آگوست نه. از آنجا که داده‌ها را باید جمع‌آوری، پالایش، بارگذاری و تست کرد، در بسیاری از شرکتها، داده‌های آگوست، حتی تا اوایل اکتبر در دسترس نیستند. راه‌حل شامل یک ماه تاخیر در مجموعه مدل است.

مرحله 5: تثبیت مسئله با داده‌ها :

تمام داده‌ها کثیف هستند. تمام داده‌ها مسئله دارند. آیا مسئله ای هست که با تکنیک های داده‌کاوی تغییر کند؟ برای مثال، برای درخت تصمیم گیری، مقادیر گمشده و چیزهای جدا، نمی‌تواند دردسر زیادی درست کند. برای بقیه، مثل شبکه‌های عصبی، می‌توانند باعث دردسرهای زیادی شوند.

متغیرهای طبقه بندی شده، با مقادیر بسیار :

متغیرهایی مثل کدپستی، کشور و کد شغل، همه از متغیرهایی هستند که اطلاعات مفیدی را بیان می‌کنند، ولی نه از راهی که بیشتر الگوریتم های داده‌کاوی می‌توانند مدیریت کنند. مقادیر ممکن بسیاری برای متغیرهایی که این اطلاعات را حمل می‌کنند وجود دارد، و مثال های کمی در داده‌های شما برای بیشتر این مقادیر وجود دارد.

متغیرهای این چنینی، یا باید گروه‌بندی شوند که در آن صورت کلاس‌های بسیار زیادی برای رابطه آنها با متغیر پیش بینی، بوجود خواهد آمد؛ یا باید با خصوصیات مطلوب از کدپستی و

غیره جایگزین شوند. جایگزینی کدپستی با کدپستی میانگین، جایگزینی شغل با میانه حقوق برای شغل و به همین صورت.

توزیع مورب و پرت، مسائلی را برای هر تکنیک داده‌کاوی که مقادیر را بطور ریاضی بکار می برد، (مثلا با ضرب وزن آنها در یکدیگر و اضافه کردن آنها به هم) ایجاد می‌کند. در بسیاری از موارد، رکوردهایی که پرت افتاده اند، خارج می‌شوند. در سایر موارد، بهتر است که مقادیر را به اندازه دامنه یکسان تقسیم کرد. گاهی هم بهترین روش برای کاهش دامنه مقادیر این متغیرها، جایگزینی هر مقدار مثلا با لگاریتم آن است.

مقادیر گم شده :

برخی از الگوریتم ها، قادرند با مقادیر گم شده نیز مانند سایر مقادیر رفتار کنند و آنها را در قوانین بیاورند. متاسفانه برخی دیگر نمی‌توانند. دور ریختن تمام این رکوردهای با مقادیر گمشده، تبعیض را معمول می‌کند چون بعید است که این قبیل رکوردها بطور تصادفی توزیع شده باشند. جایگزینی آنها با مقادیری مثل میانه‌ها و مقدار متداول‌تر، اطلاعات جعلی را می‌سازد. جایگزینی آنها با مقادیر نامطلوب، حتی می‌تواند بدتر باشد، چون الگوریتم های داده‌کاوی آنها را نمی‌توانند تشخیص دهند، مثلا جایگزینی مقدار 999 برای متغیر سن. وقتی مقادیر گمشده باید جایگزین شوند، بهترین راه‌حل این است که با ساختن یک مدل که مقادیر گمشده بعنوان متغیر هدف باشند، آنها را مستند کنیم.

مقادیری که مفاهیم آنها در طول زمان تغییر کرده است :

وقتی داده‌ها از چندین نقطه از گذشته، می‌آیند، خیلی متداول است که مفهوم برخی مقادیر برای برخی فیلدها، در طول زمان تغییر کرده باشند. ممکن است که کلاس A اعتبار، همیشه بهترین کلاس باشد، اما آن دامنه رتبه‌های اعتباری که کلاس A را تشکیل می‌دهند، از زمانی به زمان دیگر، ممکن است تغییر کند. بحث کردن درباره این، کاملا به یک انبار داده خوب طراحی شده نیاز دارد که این تغییرات در معنی بصورت یک متغیر جدیدی که می‌تواند در طول زمان معنای ثابتی داشته باشد،ثبت شده‌اند.

کدگذاری ناسازگار داده :

وقتی اطلاعات با موضوع یکسان از چندین منبع گردآوری شوند، اغلب ممکن است منابع مختلف روش های ارائه مختلفی برای داده‌ها داشته باشند. اگر این تفاوت ها دریافت نشوند، تمایزهای جعلی را اضافه می‌کنند که منجر به استنتاجات نادرستی می‌شود. با دانستن این تفاوت ها لازم است که داده‌ها به شکل رایج آنها ثبت شوند.

مرحله 6: تبدیل داده برای آوردن اطلاعات به سطح :

وقتی که داده‌ها همگذاری شدند و مسائل تثبیت شدند، هنوز داده باید برای تحلیل آماده شود. این شامل افزودن فیلدهای مشتق شده برای آوردن اطلاعات در سطح است. همچنین ممکن است این کار حذف پرت‌افتاده‌ها، گروه بندی کلاس ها برای رده بندی متغیرها، اعمال تبدیلاتی از قبیل لگاریتم‌ها و سایر کارهای مشابه باشد. آمایش داده‌ها، مسئله مهمی است که در کتابی بعنوان آمایش داده‌ها برای داده‌کاوی، بطور مفصل شرح داده شده است.

گرفتن گرایشات :

بیشتر داده‌های یکی شده، شامل سری‌های زمانی هستند. اطلاعات صورتحساب‌های عکس‌های فوری ماهانه، کارکردها، تماس ها و غیره. بیشتر الگوریتم های داده‌کاوی، سری های زمانی را درک نمی‌کنند. سیگنال‌هایی مثل "سه ماه کاهش سود سهام" نمی‌توانند نقطه به نقطه با مشاهدات هر ماه بطور مستقل، مطابقت داشته باشد. این تاحدودی به داده‌کاو بستگی دارد که اطلاعات رفتاری را با افزودن متغیرهای مشتق شده‌ای، به سطح بیاورد، مثل نسبت صرف شده در ماه‌های اخیر به ماه‌های قبلی، برای داشتن یک گرایش کوتاه مدت، و نسبت ماه‌های اخیر به همان ماه‌ها در سال قبل برای بدست آوردن یک گرایش بلندمدت.

ساختن نسبت ها و سایر ترکیبات متغیرها :

گرایشات یکی از مثال‌های آوردن اطلاعات در سطح با ترکیب متغیرهای چندگانه است. از اینها تعداد زیادی وجود دارد. اغلب این فیلدهای اضافی از روشهای موجود از طریق تحلیل‌گران دارای دانش گرفته شده‌اند، ولی کمتر توسط نرم‌افزارها توجه شده‌اند.

فیلدهای اضافه شده که روابط مهم از نظر خبره‌های آن زمینه را نشان می‌دهد، راهی است که اجازه می‌دهد که روال کنکاش از مهارتها بهره گیرد.

تبدیل شمارش ها به خصوصیات :

بسیاری از مجموعه داده‌ها، شامل مقادیر شمارش یا دلار هستند که خودشان بویژه جالب نیستند چون مطابق با سایر مقادیر، تغییر می‌کنند. خانواده‌های بزرگتر به‌نسبت خانواده‌های کوچکتر، پول بیشتری را در بقالی ها صرف می‌کنند. پول بیشتری در گوشت، محصولات، در بسته بندی بهتر، در محصولات شوینده، و همه چیز. پس مقایسه فیلد قیمت دلاری صرف شده بوسیله خانواده‌های با جمعیت متفاوت، در یک طبقه، مثل نانوایی، فقط آشکار می‌کند که خانواده‌های بزرگتر، بیشتر مصرف می‌کنند. مطلوبتر آن است که درجه مصرف هر خانواده را در هر طبقه، بررسی کنیم.

مرحله 7: ساختن مدلها :

جزئیات این مرحله از تکنیکی به تکنیک دیگر تغییر می‌کند. به عبارت کلی، این مرحله‌ای است که بیشتر کارهای ساخت یک مدل در آن اتفاق می‌افتد. در داده‌کاوی جهتدار، مجموعه آموزشی برای تولید تفسیری از متغیرهای مستقل یا هدف بر اساس متغیرهای وابسته یا ورودی، استفاده می‌شود. این تفسیر ممکن است یک شکل شبکه عصبی را بگیرد، یا درخت تصمیم‌گیری، یا یک گراف پیوند، یا سایر ارائه‌های دیگر از رابطه بین هدف و سایر فیلدهای پایگاه داده. در داده‌کاوی غیرجهتدار، هیچ متغیر هدفی وجود ندارد. مدل، روابط بین رکوردها را می‌یابد و آنها را بصورت قوانین وابستگی یا با تخصیص آنها به خوشه‌های متداول، بیان می‌کند.

ساختن مدل ها، اولین گام از روال داده‌کاوی است که بدرستی با نرم‌افزارهای داده‌کاوی مدرن بصورت اتوماتیک درآمده است. به همان دلیل، زمان کمتری را در پروژه داده‌کاوی، می‌گیرد.

مرحله 8: ارزیابی مدل ها :

این مرحله تعیین می‌کند که مدل ها کار می‌کنند یا نه. ارزیابی مدل ها باید به سوالات از این قبیل، پاسخ دهد:

1- چقدر یک مدل دقیق و قابل درک است؟
2- مدل چقدر توانسته داده‌های مشاهده شده را تشریح کند؟
3- چه میزان اطمینان از پیش بینی های مدل وجود دارد؟

البته پاسخ به این سوالات به نوع مدل ساخته شده بستگی دارد.

ارزیابی مدل های توصیفی :

قانون، "اگر (ایالت = 'MA') آنگاه منبع گرما نفت است"، از قانون، "اگر(ناحیه = 339 یا 351 یا 413 یا 508 یا 617) باشد آنگاه منبع گرمایش نفت خواهد بود"، بسیار توصیفی‌تر است. حتی اگر دو قانون معادل هم باشند، اولی به نظر می‌رسد که توصیفی‌تر است.

قدرت بیان ممکن است بطور ذهنی به نظر برسد، ولی در اینجا در حقیقت یک راه تئوری برای اندازه‌گیری آن وجود دارد، که کوچکترین طول توصیف (MDL) نامیده می‌شود. برای یک مدل، MDL، عبارت است از تعداد بیت‌هایی که آن مدل برای کدگذاری قانون و تمام استثناءهای آن، می‌گیرد. قانونی بهتر است که به بیت‌های کمتری نیاز دارد. برخی ابزارهای داده‌کاوی MDL را استفاده می‌کنند تا ببینند که کدام مجموعه قانون را نگه دارند و کدام را دور بریزند.

ارزیابی مدل های جهتدار :

مدل های جهتدار در دقتشان در داده‌های نادیده اولیه، سنجیده می‌شوند. ارزیابی هر مدلی به زمینه آن بستگی دارد. مدل های یکسان می‌توانند برطبق یک پیمانه، بد و در دیگری خوب باشند. محققان یک هدف از ساخت مدل ها دارند و آن امکان درک تمامیت مسئله است.

ارزیابی مدل می‌تواند در سطح کل مدل باشد یا در سطح پیشگوهای تکی. دو مدل با دقتهای سراسری یکسان، ممکن است سطوح مختلفی از ناسازگاری را در پیشگوهای تکی شان داشته

باشند. یک درخت تصمیم‌گیری، برای نمونه، یک نرخ خطای دسته‌بندی سراسری دارد، ولی هر شاخه و برگ درخت نرخ خطای خودش را دارد.

ارزیابی طبقه‌بندها و پیشگوها :

برای وظایف پیشگویی و دسته‌بندی، دقت با عبارت نرخ خطا، سنجیده می‌شود. نرخ خطای دسته‌بندی روی یک مجموعه آزمایشی از پیش دسته‌بندی شده، بکار می رود برای تخمین نرخ خطا هنگامی که رکوردهای جدید دسته‌بندی می‌شوند. البته این روال فقط هنگامی معتبر است که مجموعه آزمایشی در نسبت بزرگتری ارائه شوند.

ارزیابی تخمین‌گرها :

برای وظایف تخمین، دقت می‌تواند با عبارت تفاوت بین نمره پیش بینی شده و اندازه واقعی، بیان شود. دقت هر جزئی که تخمین زده می‌شود و دقت کل مدل، هر دو مطلوب و مورد نظر ماست. ممکن است مدلی برای یک دامنه از مقادیر ورودی، کاملا صحیح باشد و برای دامنه دیگری نباشد. روش استاندارد برای توصیف دقت مدل تخمین، اندازه گیری میزان دوری تخمین از میانگین. ولی می‌توان بسادگی با کم کردن مقدار تخمین از مقدار واقعی یا با میانه گیری برای مقادیر بی معنی، آنرا بدست آورد."واریانس کوچکتر بمعنی دقت بالاتر تخمین است."

مقایسه مدل ها با استفاده از lift :

مدل های جهتدار، که با شبکه‌های عصبی، الگوریتم ژنتیک و درخت تصمیم همگی برای به انجام رساندن برخی وظایف ساخته شده‌اند. چرا نباید آنها درباره توانایی هایشان برای دسته‌بندی، تخمین و پیش‌بینی، مقایسه شوند؟ یک روش بسیار معمول برای مقایسه کارایی دسته‌بندی مدل ها،استفاده از نسبتی بنام Lift است. این مقیاس می‌تواند برای مقایسه مدل های طراحی شده برای سایر وظایف نیز بخوبی تطبیق داده شود. درحقیقت lift تغییر تمرکز از یک کلاس خاص را وقتی

مدل برای انتخاب یک گروه از جمعیت کل، استفاده می‌شود، اندازه می گیرد.

$$lift = P(class_t| sample) / P(class_t | population)$$

فرض کنید می خواهیم مدلی را بسازیم برای پیش بینی اینکه چه کسی تمایل به پاسخ به اشتراک پست مستقیم دارد. معمولا، ما مدلی می سازیم با استفاده از یک مجموعه داده آموزشی پیش طبقه‌بندی شده، و اگر لازم شد، یک مجموعه اعتبارسنجی پیش طبقه‌بندی شده. حالا ما آماده ایم که مجموعه تست را برای محاسبه lift مدل بکار بریم.

دسته‌بندی کننده، رکوردها را در مجموعه تست بعنوان "پیش بینی شده برای پاسخ" یا "پیش‌بینی نشده برای پاسخ" دسته‌بندی می‌کند. البته این همیشه درست نیست، ولی اگر مدل به هیچ وجه مزیتی نداشته باشد، گروه علامت گذاری شده بعنوان "پیش بینی شده برای پاسخ"، شامل یک نسبت بالاتر از پاسخ دهندگان واقعی است به نسبت کل مجموعه تست. به این رکوردها توجه کنید. اگر مجموعه تست شامل 5 درصد از پاسخ دهندگان واقعی باشد و نمونه، شامل 50 درصد از پاسخ دهندگان واقعی باشد، lift مدل 10 است.(50/5)

آیا مدلی که Lift بیشتری داشته باشد، بهتر است؟ البته فهرستی از افراد، که نیمی از کل آنها پاسخ دهنده خواهند بود، به فهرستی که فقط یک چهارم آنها پاسخ دهنده‌اند، ترجیح داده می‌شود. درسته؟ نه لزوما- ممکن است که آن لیست فقط شامل 10 نام از پاسخ دهندگان باشد.

نکته اینجاست که lift یک تابع از اندازه نمونه است. اگر یک دسته بند فقط 10 تا از پاسخ دهندگان را بردارد و این 100 درصد زمان باشد، به مقدار20 =lift خواهد رسید، بیشترین مقدار lift ممکن وقتی که جمعیت حاوی 5 درصد پاسخ دهندگان باشد.

مرحله 9: استقرار مدل ها :

استقرار مدل به معنی بردن آن از محیط داده‌کاوی به محیط رتبه بندی است. فرایند ممکن است ساده یا سخت باشد. در بدترین حالت(که در بیش از یک شرکت دیده شده) مدل در یک محیط مدلسازی خاص با استفاده از نرم افزاری که هیچ جایی اجرا نمی‌شود، توسعه داده می‌شود.

یک مسئله شایع دیگر این است که مدل از متغیرهای ورودی‌ای استفاده کرده باشد که در داده واقعی وجود ندارد. این مسئله نباید وجود داشته باشد از آنجا که متغیرهای ورودی مدل از فیلدهای برگرفته از داده اصلی در مجموعه مدل، گرفته شده‌اند.

مرحله 10: ارزیابی نتایج

مرحله 11: شروع دوباره :

از هر پروژه داده‌کاوی پرسش‌هایی بسیاری برمی‌خیزد. این چیز بسیار خوبی است. این به این معنی است که روابط جدیدی اکنون دیده می‌شوند که قبلا دیده نمی‌شدند. روابط جدید کشف شده، فرضیه‌های جدیدی را پیشنهاد می‌کنند که باید آزمایش شوند و بنابراین روال داده‌کاوی از ابتدا آغاز می‌شود.

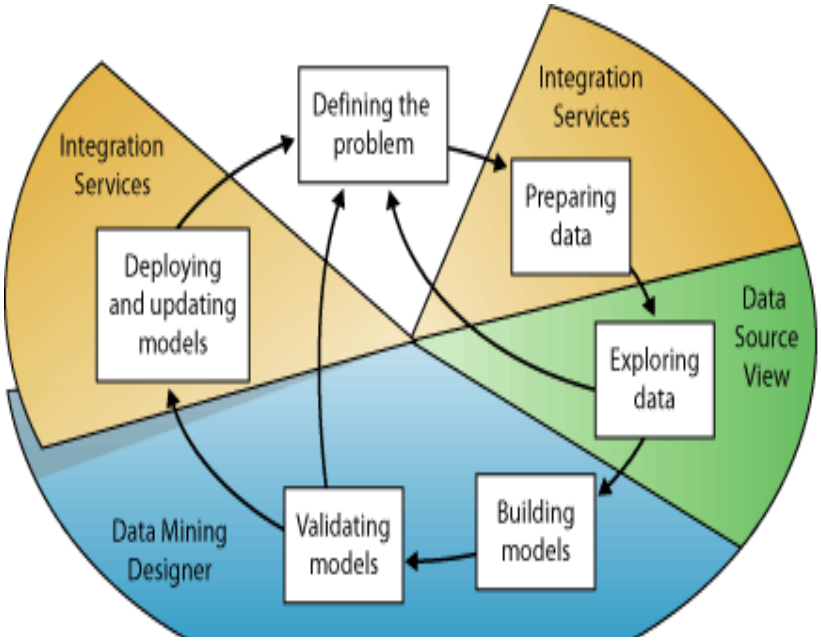
۲-٤-۱ دیگر نظریه ها (مراحل داده کاوی) :

1- تعیین اطلاعات گذسته
2- تمیز دادن و پردازش های اولیه . در این مرحله خطاهای داده تصحیح می شوند و داده های اشتباه جایگزین می شوند . این مرحله ممکن است تا 60 درصد از زمان داده کاوی را به خود اختصاص بدهد
3- یکپارچه سازی داده ها. معمولا داده ها از منابع متفاوتی جمع آوری می شوند و باید به صورتی در آیند که یک مخزن داده یی مناسب را ایجاد کنند تا بتوان عملیات داده کاوی را بهتر انجام داد
4- انتخاب مجموعه ای از داده های هدف
5- یافتن ویژگی های مورد استفاده
6- نمایش داده ها به صورتی که بتوان برای داده کاوی مورد استفاده قرار گیرند
7- انتخاب عمیات داده کاوی (طبقه بندی شده،خوشه بندی و . . .)
8- انتخاب روش داده کاوی(شبکه های عصبی، درخت تصمیم و . . .)
9- داده کاوی برای یافتن الگوی مناسب
10- ارزیابی و تحلیل الگوی بدست آمده و حذف الگوی نا مناسب
11- تفسیر نتایج داده ها و استنتاج ار اطلا عات با ارزش

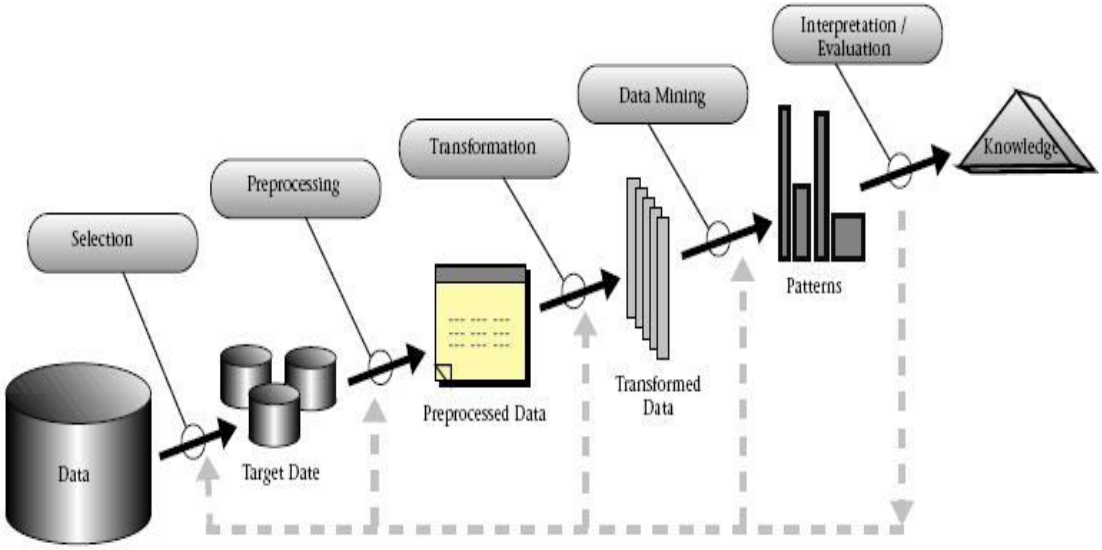
باید توجه داشت که جمع آوری و حفاظت از داده ها نکته بسیار مهمی است. اصولا چون قالب و نوع داده ها در طول زمان تغیر می کند ممکن است بسیاری از داده های موجود در قالب های متفاوتی با شد و هم چنین به بسیاری از داده های قدیمی از بین رفته و دور ریخته می شود در حالی که ممکن است اهمیت این داده ها از داده های جدید به هیچ وجه کمتر نباشد. هم چنین

بـه علـت اینکـه داده هـا میتواننـد از منـابع متلفـی داخلـی و خـارجی ماننـد: کارکنـان شـرکت، مـدیران شـرکت، مشـتریان، کـار فرمایـان و کارکنـان باشـند بـازهم ممکـن اسـت قالـب داده هـا بـا هـم یکسـان نباشند. بـه همـین دلیـل انتخـاب داده هـای درسـت و یکپارچـه سـازی قالـب آن هـا بـه منظـور اسـتفاده در داده کـاوی از اهمیـت بـالایی بـر خـور دار اسـت. در شـکل 2 مـی تـوان مراحـل داده کـاوی را بـه اختصار مشاهدا کرد

1- درک قلمرو یا بیان مسئله و فرموله کردن فرضیه
2- انتخاب و جمع آوری داده ها
3- تبدیل داده ها
4- کاوش در داده ها
5- تفسیر نتیجه یا تفسیر مدل و رسیدن به نتایج



مراحل داده کاوی عموم-شکل 1



مراحل داده کاوی آکادمیک- شکل2

۵-۲ وظایف داده‌کاوی.Error! Reference source not found :

داده‌کاوی می تواند برای حل صدها مسئله کسب و کار بکار آید. بر اساس طبیعت این مسائل، ما آنها را به وظایف داده‌کاوی زیر دسته‌بندی می‌کنیم.

1- دسته‌بندی :

دسته‌بندی یکی از رایج‌ترین وظایف داده‌کاوی است. مسائل تجاری‌ای مانند تحلیل رویگردانی، مدیریت ریسک و هدف‌گیری موردی، شامل دسته‌بندی می‌شوند.

دسته‌بندی به تخصیص موارد به گروه ها براساس یک خصوصیت قابل پیش‌بینی، برمی‌گردد. هر مورد شامل مجموعه ای از خصوصیات است که یکی از آنها خصوصیت کلاس(خصوصیت قابل پیش‌بینی) نام دارد. این وظیفه شامل یافتن مدلی است که خصوصیت کلاس را براساس تابعی از خصوصیات ورودی تشریح کند. برای آموزش مدل رده‌بندی، شما به مقدار کلاس موارد ورودی در مجموعه داده‌های آموزشی نیاز دارید. الگوریتم‌های داده‌کاوی که به یک هدف برای یادگیری نیاز دارند، الگوریتم‌های نظارت شده، نام دارند.

الگوریتم‌های رایج دسته‌بندی شامل درخت‌های تصمیم، شبکه عصبی، و ناوی بیز است.

2- خوشه‌بندی :

خوشه‌بندی که گاهی قطعه‌بندی نیز خوانده می‌شود برای شناسایی گروه‌بندی طبیعی موارد بر اساس یک مجموعه خصوصیت بکار می‌رود. موارد درون یک گروه بیشترین یا کمترین میزان شباهت را در مقادیر خصوصیات دارند.

الگوریتم خوشه بندی مجموعه داده را براساس این دو خصوصیت به سه گروه تقسیم کرده است.

خوشه‌بندی یک وظیفه داده‌کاوی بدون نظارت است. هیچ خصوصیت تکی برای هدایت روال آموزش بکار نمی‌رود و تمام خصوصیات ورودی بطور یکسان هدف هستند. بیشتر الگوریتم‌های خوشه‌بندی، مدل را در تعداد تکرارهای معینی می‌سازند و هنگامی که مدل همگرا شد، متوقف می‌شوند و ان هنگامی است که مرزهای گروه‌ها تثبیت شد.

3- تخمین:

تخمین با نتایجی که با ارقام پیوسته نشان داده شده‌اند سروکار دارد. در تخمین داده‌های ورودی داده می‌شوند و به رقمی در متغیرهای ناشناس مداوم چون درآمد یا تراز کارت اعتباری ختم می‌شود. در عمل، تخمین اغلب برای انجام دسته بندی استفاده میشود.

یک شرکت کارتهای اعتباری که مایل است یک فضای تبلیغاتی را در پاکتهای صورتحساب به یک تولید کننده پوتین اسکی بفروشد، باید مدل دسته بندی تهیه کند که همه دارندگان کارتها را در یک یا دو دسته قرار دهد: اسکی باز یا غیر اسکی باز.

روش دیگر ایجاد مدلی است که به هر دارنده کارت، یک امتیاز تمایل به اسکی تخصیص می دهد. این ارقام می تواند صفر و یک باشد که نشانگر احتمال تخمین زده شده است که دارنده کارت یک اسکی باز است یا نه. عمل دسته‌بندی اکنون به ایجاد امتیازی آستانه ای منجر می‌گردد. هر کسی که امتیازی بیشتر یا مساوی با امتیاز آستانه داشته باشد به عنوان اسکی باز قلمداد می‌شود. هر کسی که امتیازی کمتر از امتیاز مورد نظر را داشته باشد اسکی باز محسوب نمی

گردد. شرکت تولید پوتین‌های اسکی برای پست پانصد هزار اوراق تبلیغاتی بودجه‌ریزی نموده است. اگر از روش دسته‌بندی استفاده شده و یک و نیم میلیون نفر اسکی باز تعیین شده‌اند پس به راحتی می توان به صورت تصادفی تبلیغات را در صورتحسابهای پانصد هزار نفر منتخب از آن افراد قرارداد. اگر از طرف دیگر هر دارنده کارت، امتیاز تمایل به اسکی را داشته باشد می‌توان تبلیغات را برای پانصد هزار از محتمل‌ترین کاندیداها فرستاد.

مثالهایی دیگر از تخمین:

1- تخمین تعداد فرزندان در یک خانواده
2- تخمین درآمد کل یک خانواده
3- تخمین عمر یک مشتری
4- تخمین احتمال اینکه فردی به یک تقاضای تغییر میزان اعتبار بانکی پاسخ دهد.

4- وابستگی:

وابستگی یکی دیگر از وظایف معمول داده‌کاوی است که تحلیل سبد بازار نیز خوانده می‌شود. یک مسئله کسب و کار معمول وابستگی، تحلیل یک جدول تراکنش فروش و شناسایی آن محصولاتی است که اغلب در یک سبد خرید یکسان قرار می گیرند. استفاده رایج از وابستگی شناسایی مجموعه رایج اقلام خرید(مجموعه اقلام متناوب) و قوانینی برای فروش-متقاطع است.

در عبارات وابستگی، هر محصولی، یا بطور کلی هر جفت خصوصیت/مقداری، یک قلم در نظر گرفته می‌شود. وظیفه وابستگی دو هدف دارد: یافتن مجموعه اقلام متناوب و یافتن قوانین وابستگی.

بیشتر الگوریتم‌های وابستگی برای یافتن مجموعه اقلام متناوب، مجموعه داده‌ها را چندین بار پیمایش می‌کنند. آستانه تناوب (پشتیبانی)، قبل از شروع پردازش، توسط کاربر تعیین می‌شود. برای مثال پشتیبانی برابر با 2 درصد به معنی است که مدل فقط آن اقلامی که در 2 درصد از کارت‌های خرید دیده می‌شوند را تحلیل کند. یک مجموعه قلم می‌تواند مانند مجموعه {محصول="پپسی"، محصول="چیپس"، محصول="آبمیوه"} دیده شود. هر مجموعه قلم یک اندازه دارد که تعداد آیتم‌های آن است. مثلا اندازه این مجموعه آیتم برابر با 3 است.

جـدای از شناسـایی مجموعـه اقـلام برمبنـای پشـتیبانی، بیشـتر الگـوریتم‌هـای وابسـتگی، قـوانین را نیـز می‌یابنـد. یـک قـانون وابسـتگی شـکلی ماننـد A,B =>C بـا یـک احتمـال دارد، کـه A و B و C همگـی مجموعـه اقـلام هسـتند. همچنـین احتمـال در ادبیـات داده‌کـاوی اعتمـاد نیـز خوانـده مـی‌شـود. احتمـال یک مقدار آستانه است کـه یـک کـاربر قبـل از آمـوزش یـک مـدل وابسـتگی، نیـاز دارد کـه تعیـین کنـد. برای مثـال، ایـن یـک قـانون معمـول اسـت: محصـول= "پپسـی" و محصـول= "چیـپس" => محصـول= "آبمیوه" با احتمال 80%. تفسـیر ایـن قـانون بسـیار سرراسـت اسـت. اگـر یـک مشـتری پپسـی و چیـپس خرید، با احتمال 80 درصد ممکن است که او آبمیوه نیز بخرد.

5- رگرسیون :

وظیفه رگرسـیون مشـابه بـا دسـته‌بنـدی اسـت. تفـاوت اساسـی در خصـویت پـیش‌بینـی اسـت کـه یـک عـدد پیوسـته اسـت. تکنیـک رگرسـیون سـالها در حـوزه آمـار مطالعـه شـده اسـت. رگرسـون خطـی و منطقی از روش‌هـای بسـیار رایـج رگرسـیون هسـتند. سـایر تکنیـک‌هـای رگرسـیون شـامل درخـت‌هـای رگرسیون و شبکه‌های عصبی است.

وظیفه رگرسیون مـی توانـد بیسـاری از مسـائل کسـب و کـار را حـل کنـد. بـرای مثـال، آنهـا مـی تواننـد در پـیش‌بینـی نـرخ‌هـای خریـداری و آزادسـازی کـوپن بـر اسـاس ارزش وجـه، روش توزیـع و حجـم توزیع یا پیش‌بینی سرعت باد براساس دما، فشار هوا و رطوبت، بکار روند.

6- پیشگویی :

پیشگویی یـک وظیفـه داده‌کـاوی مهـم دیگـری اسـت. ارزش سـهام MSFT فـردا چگونـه خواهـد بـود؟ مقدار فـروش پپسـی در مـاه آینـده چگونـه خواهـد بـود؟ پیشـگویی مـی توانـد بـه ایـن سـوالات پاسـخ دهد.

در عمل پیش‌بینی، تنهـا روش بـرای بررسـی صـحت مـدل، دیـدن آینـده و مقایسـه نتیجـه مـدل و پدیـده واقع شـده مـی‌باشـد. هـر یـک از تکنیکهـای اسـتفاده شـده در دسـته بنـدی و تخمـین را مـی‌تـوان بـرای استفاده در پـیش بینـی تطبیـق داد. از داده‌هـای پیشـین بـرای تهیـه یـک مـدل کـه بیـانگر رفتـار مشـاهده

کنونی است استفاده می‌شود. وقتی این مدل برای ورودی های کنونی به کار رفت نتیجه کار پیش‌بینی رفتار آینده خواهد بود.

7- تحلیل توالی :

تحلیل توالی برای یافتن الگوها در رشته‌های گسسته بکار می‌رود. یک توالی از مقادیر(یا نواحی) گسسته ترکیب شده است. برای مثال، یک توالی DNA یک رشته مرکب از چهار ناحیه مختلف A، G، C و T است. یک توالی کلیک وب شامل رشته‌هایی از URLهاست. خریدهای کامپیوتری را هم می توان بصورت داده‌های توالی مدل کرد. برای مثال، یک مشتری ابتدا یک کامپیوتر، سپس اسپیکرها، و در نهایت یک دوربین وبی می‌خرد. تفاوت داده‌های توالی و رشته‌های زمانی در این است که رشته‌های زمانی شامل اعداد پیوسته و توالی شامل مقادیر گسسته است.

داده‌های توالی و وابستگی در اینکه هر یک شامل یک مجموعه آیتم یا ناحیه هستند، شبیه می باشند. تفاوت آنها در این است که مدل‌های توالی انتقال‌های نواحی را تحلیل می‌کنند و مدل وابستگی فرض می کند که هر فقره در یک کارت خرید برابر یا مستقل باشد. با مدل توالی، خرید یک کامپیوتر قبل از اسپیکرها با خرید اسپیکرها قبل از کامپیوتر متفاوت است. درحالی که با یک الگوریتم وابستگی این دو توالی خرید، مجموعه اقلام یکسانی را می‌سازند.

تحلیل توالی، نسبتا یک وظیفه داده‌کاوی جدید است. و در دو نوع کاربرد بسیار مهم شده است: تحلیل ثبت‌های وب و تحلیل DNA. امروزه تکنیک‌های توالی مختلفی مانند زنجیره مارکوف در دسترس ما هستند. محققان در این زمینه بطور فعالی در حال تحقیق هستند

8- تحلیل انحراف:

تحلیل انحراف برای یافتن موارد نادری است که با دیگران بسیار متفاوت هستند. این که همچنین کشف outlier نیز نامیده می‌شود که تشخیص تغییرات معنی‌دار از رفتار مشاهده شده قبلی است. تحلیل انحراف می‌تواند در کاربردهای بسیاری استفاده شود. رایج‌ترین استفاده از آن در کشف

تقلبات کارتهای اعتباری است. شناسایی موارد غیرعادی از بین میلیونها تراکنش یک چالش بسیار مشکل است. سایر کاربردها می تواند شامل کشف ورود سرزده به شبکه، تحلیل خطای تولید و غیره باشد.

هیچ استانداردی برای تحلیل انحراف وجود ندارد و این هنوز یک موضوع باز برای تحقیقات است. اما تحلیلگران معمولا نسخههای اصلاح شده الگوریتمهای درختهای تصمیم، خوشهبندی، یا شبکههای عصبی را برای این کار بکار میگیرند. بمنظور ساختن قوانین معنیدار، تحلیلگران نیاز به داشتن نمونههای بسیار از این موارد نادر در مجموعههای آموزشی خود دارند.

9- نمایهسازی:

گاهی اوقات هدف دادهکاوی تنها توصیف آن چیزی است که در یک پایگاه داده پیچیده درجریان است. نتایج نمایهسازی درک ما را از مردم، محصولات یا فرآیندهایی که دادهها را در مرحله اول تولید کردهاند افزایش میدهد. توصیف خوب رفتار اغلب توضیح خوبی هم به همراه دارد.

شکاف جنسیتی مشهور در سیاست آمریکا مثالی از این دست است که چگونه این توصیف ساده که تعداد زنان حامی حزب دمکرات بیش از مردان است می تواند توجه بیشتر و مطالعات تکمیلی را برای روزنامهنگاران، جامعهشناسان، اقتصاددانان و دانشمندان علوم سیاسی ایجاد کند.

درخت های تصمیم ابزار مفیدی برای نمایه سازی می باشد. قوانین وابستگی و خوشه بندی را نیز می توان برای نمایه سازی ها استفاده نمود.

سه وظیفه نخست مثال هایی از دادهکاوی جهتدار است. در دادهکاوی جهت دار همیشه یک متغیر هدف وجود دارد برخی مواقع دستهبندی میشود، تخمین زده میشود یا پیش بینی میشود. روال ساختن یک دستهبند، با یک مجموعه پیش تعریف شده کلاس ها و مثالهایی از رکوردهایی شروع میشود که در حال حاضر بخوبی کلاس بندی شده اند. بطور مشابه، روال ساختن یک تخمین گر، با دادههای سابقه ای شروع میشود که مقدار متغیر مقصد برای آنها در حال حاضر

مشخص است. وظیفه مدل سازی یافتن قوانینی است که مقادیر شناخته شده متغیر هدف را توضیح دهد.

در داده‌کاوی غیر جهتدار، هیچ متغیر هدفی وجود ندارد. گروه بندی وابستگی و خوشه بندی، داده‌کاوی بدون جهت هستند. وظیفه داده‌کاوی یافتن الگوهای موازی است که به هیچ متغیری مقید نیستند. یک شکل رایج داده‌کاوی غیرجهتدار، خوشه بندی است، که گروه‌هایی از رکوردهای مشابه را می یابد بدون هیچ دستورالعملی درباره اینکه به چه متغیری باید بیشتر توجه شود. داده‌کاوی غیر جهتدار با طبیعتش توصیف شده است.

بیان مسئله و فرموله کردن فرضیه :

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 داده را جهت کاوش آماده نموده یا نتایج را به طور صحیح تفسیر نمود. برای استفاده بهتر از داده کاوی باید یک بیان واضح از هدف داشت. در این مرحله انچه نیاز است ترکیبی از تخصص یک زمینه کاربردی و یک مدل داده کاوی است و شاید بتوان گفت یک تقابل نزدیک سر یک مسئله واحد و چندین فرضیه فرموله شده بین متخصصین داده کاوی و متخصصین کاربردی میباشد.

1- انتخاب و جمع آوری داده ها:

این مرحله درارتباط با چگونگی تولید و جمع آوری داده ها است.

بطور کلی، دو امکان وجود دارد:

روش آزمون طراحی: زمانی است که فرایند تولید داده ها تحت کنترل یک متخصص کاربردی(مدل ساز سیستم) باشد.

روش دیداری: امکان دوم زمانی مطرح است که متخصص قادر به تولید فرآیند نیست یعنی تولید داده بصورت تصادفی در نظر گرفته شود.

پس از اینکه داده ها جمع اوری شدند یا در فرایند جمع اوری داده ها تا اندازه ای قرار گرفتند، توزیع نمونه گیری کاملا نامعلوم است.(یعنی داده هایی که بعدا برای تست و بکارگیری آن مدل بکار می روند از چند نمونه مشابه استفاده می شوند.)

نکته: برای فرایند داده کاوی داده ها ی مورد نیاز موجود در انبار داده ها باید انتخاب شوند. درک این مطلب که برای ارزیابی یک مدل که بعدا برای تست و بکارگیری آن مدل بکار می رود، موفقیت آمیز باشد، بسیار مهم است در غیر اینصورت نتایج درستی حاصل نمی گردد.

مثلا انبار داده ها شامل انواع مختلف و گوناگونی از داده ها است به عنوان مثال در یک پایگاه داده های مربوط به سیستم فروشگاهی، اطلاعاتی در مورد خرید مشتریان، خصوصیات آماری آنها، dispatcherها (توزیع کنندگان)، مشتریان، حسابداری و ...وجود دارند که همه آنها در داده کاوی مورد نیاز نیستند.

2- پیش پردازش ها یا تبدیل داده ها

زمانی که که داده های مورد نیاز از پایگاه داده های موجود در انبار داده ها "جمع اوری "شدند و داده های مورد کاوش مشخص گردیدند، معمولا به تبدیلات خاصی روی داده ها نیاز است که شامل حداقل دو مرحله متداول می باشد:

آشکارسازی (حذف) داده های غیرعادی:

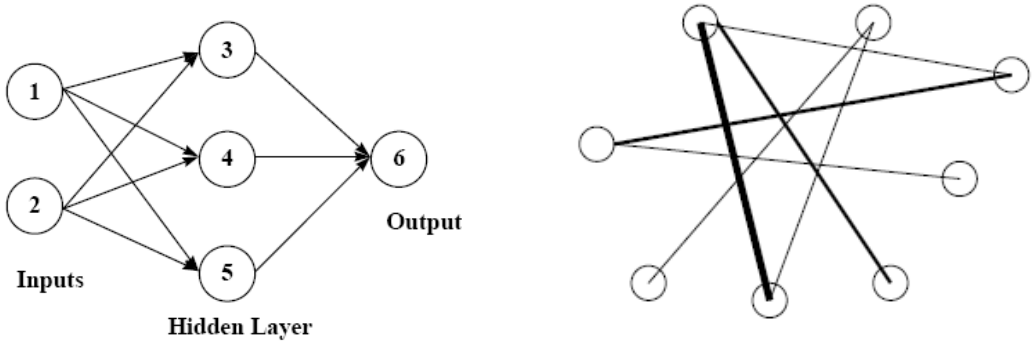
داده های غیرعادی یا غیر معمول درحقیقت داده های نتیجه سنجش خطاها، کدنویسی و ثبت خطاها است. دراینجا باید یا .1داده های غیرعادی را تشخیص داد و خذف کرد ویا .2باید روش های قوی مدل سازی رابگونه ای توسعه داد که نسبت به این نوع داده ها غیر حساس باشند.

ویژگی های مقیاس بندی، رمزگذاری و انتخاب:

در تبدیل داده ها توصیه میشود که داده ها را جهت تحلیل و بررسی مقیاس بندی و ورمزگذاری کرد .مثلا یک مشخصه با دامنه [0و1] ودیگری با دامنه [1000و100-] دارای ارزش مشابهی در تکنیک های اعلام شده نیستند. که در صورت نادیده گرفتن همین تفاوت در دامنه داده ها، روی نتایج نهایی داده کاوی تاثیر خواهند گذاشت.

3- برآورد مدل یا کاوش در داده ها

در ایـن مرحلـه داده هـای تبـدیل شـده بـا اسـتفاده از تکنیکهـا و عملیاتهـای داده کـاوی مـورد کـاوش قـرار مـی گیرنـد تـا الگوهـای مـورد نظـر کشـف شـوند. یـا بـه عبـارتی دیگـه، انتخـاب و پیـاده سـازی تکنیـک هـای داده کـاوی در ایـن مرحلـه صـورت میگیـرد. لبتـه ایـن فراینـد خیلـی روشـن و واضـح نیست زیرا هنگام پیاده سـازی ممکـن اسـت کـه مبتنـی بـر چنـدین مـدل در یـک فرآینـد تکراریباشـد. (این مدل ها بطور کامـل تـر در مباحـث مربـوط بـه مفـاهیم انـواع دسـته بنـدی،درختان تصـمیم و قـوانین تصمیم، شبکه های عصبی، انواع الگوریتم ها و... پیاده سازی می شوند)



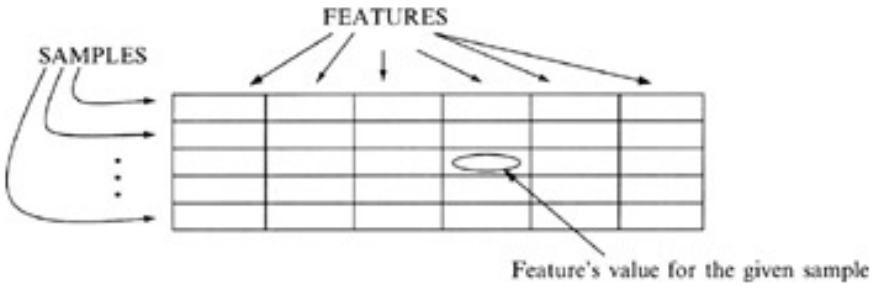
یک شبکه عصبی - شکل 3

4- تفسیر نتیجه یا تفسیر مدل و رسیدن به نتایج

اطلاعـات اسـتخراج شـده بـا توجـه بـه هـدف کـاربر تجزیـه و تحلیـل شـده و بهتـرین نتـایج بایـد در تصمیم گیـری کـاربر مـوثر مـی باشـند .هـدف از ایـن مرحلـه تنهـا ارائـه نتیجـه (بصـورت منطقـی و یـا نموداری) نیست، بلکه پالایش اطلاعات ارایه شده به کاربر نیز از اهداف مهم این مرحله است .

۶-۲ آماده سازی داده ها

مدل استاندارد داده ها:



جدول داده خام -شکل4

مدل استاندارد داده های ساخت یافته که حاوی فیلدهایی با مقادیر عددی می باشند برای داده کاوی دارای حالات مختلفی می باشند. در داده کاوی جهت نمایش اشیای ذخیره شده "خصیصه "یا "ویژگی "و جهت نمایش رکوردهای سطرها، از اصطلاح "نمونه "یا حالت " استفاده می شود. در آماده سازی داده ها گاهی اوقات فقط به عنوان یک مرحله از فرایند داده کاوی تلقی می شود.

در اماده سازی داده ها دو وظیفه داریم:

1- ساماندهی داده ها به یک شکل استاندارد که برای پردازش بوسیله تکنیک های داده کاوی آماده شوند(یک شکل استاندارد، یک جدول رابطه ای است).
2- آماده سازی مجموعه داده هایی که به بهترین عملکرد داده کاوی منجر شود.

تبدیل و تغییر وضعیت داده های خام

برای تبدیل یا تغییر وضعیت داده های خام ما باید از چند روش یا تکنیک تغییر داده ها (البته بسته به نوع داده ها) یکی را انتخاب کنیم.

* نرمال سازی

هدف از نرمالسازی حذف افزونگی داده و باقی نگهداشتن وابستگی بین داده های مرتبط از طریق ایجاد رابطه ستون های غیر کلیدی در هر جدول کلید است. این فرآیند اغلب باعث

ایجـاد جـداول بیشـتر مـی شـود ولـی از ایـن طریـق انـدازه پایگـاه داده را کـاهش داده و بهبـود کارائی داده را تضمین می کند.

با توجه به وضعیت ممکن است داده ها از چند پایگاه داده نرمال شده استخراج شوند و در یک انبار داده غیر نرمال قرار گیرد .این روش برای مخزن داده Data warehouse استاندارد خوبی است.

BIG UGLY TABLE										
SaleNo	SaleDate	ProductNo	Qty	Amount	Salesrep	CustomerNo	First	Last	Address	CreditLimit
12345	Aug 12 2002	AQX88916	1	23.95	Dave Williams	4649-4673	Richard	Johnston	14 West Avenue	1000
12346	Aug 12 2002	AQX88916	7	167.65	Sara Thompson	1113-7741	Wayne	Jones	42 York Street	<null>
12347	Aug 13 2002	AHL46785	3705	5001.75	Li Qing	1166-3461	Amelia	Waverley	995 Forth Street	<null>
12348	Aug 13 2002	DHU69863	50	118.5	Sara Thompson	<null>	<null>	<null>	<null>	<null>
12349	Aug 14 2002	DHU69863	940	2227.8	Sara Thompson	1166-3461	Amelia	Waverley	995 Forth Street	<null>
12350	Aug 14 2002	DHU69863	42	99.54	Sara Thompson	7671-3496	Antonio	Gonzales	55B Granary Lane	<null>
12351	Aug 14 2002	AQX88916	55	1317.25	Dave Williams	6794-1674	Diane	Adams	364 East Road	1500

نمونه جدول استخراج داده- شکل 5

مقیاس دهی اعشار: .مقیاس دهـی اعشـار نقطـه اعشـاری را انتقـال مـی دهـد امـا بیشـترین مقـادیر اصـلی را حفظ می کند .مقیاس کلی و انتخابی، مقادیر را در دامنه1- تا1 برقرار میکند.

$$V'(i) = V(i)/10^K$$

معادله فوق به ازای کوچکترین مقدار kبا این فرض که1 >(|V'(i)|)باشد، تعریف شده است.

نرمال سازی حداقل:حـداکثر .محاسـبه خودکـار مقـادیر حـداقل و حـداکثر نیازمنـد جسـت جـوی بیشـتر در میـان داده هـا مـی باشـد کـه ارزیـابی و بـراورد از ایـن مقـادیر ممکـن اسـت کـه موجـب انباشـتگی مقادیر نرمال شده گردد. (10و20و30) که حاصل در باز [0و1] بدست می آید.

$$V'(i) = \frac{(V(i) - \min(V(i)))}{(\max(v(i)))}$$

نرمال سازی انحـراف معیـار: نرمـال سـازی بـه روش انحـراف معیـار در اغلـب مـوارد بـا انـدازه گیـری فاصله بین بازه ها بخوبی کار میکند.

بـرای یـک مـورد i، مقـدار مشـخص بـا اسـتفاده از معادلـه زیـر تبـدیل مـی شـود:بـرای مجموعـه V={1و2و3} ، باشدmean(V)=2،Sd(V)=1 می باشد و مجموعه مقادیر نرمال شده

V'={-1و0و1} خواهد بود.

$$V'(i) = \frac{(V(i) - mean(V))}{sd(V)} \quad Sd = \sqrt{S^2} \quad S^2 = \frac{\sum (x_i - x')^2}{n-1} \quad x' = \frac{\sum_{xi}}{n}$$

*یکنواخت سازی داده ها

یک خصیصه عددی مانند y ممکن است که بر روی مقادیر مختلفی بصورت متفاوت عمل کند این نکته را نیز باید درنظر داشت که برای بسیاری از تکنیک های داده کاوی تفاوتی بین این مقادیر با اهمیت نیستند که ممکن است با تغییر در آنها موجب کاهش ان شوند .با این وجود گاهی مواقع با یکنواخت کردن مقادیر معتبر، منجر به کاهش پیچیدگی نتایج حاصله گردد.

برای مثال یکی از روش های یکنواخت سازی داده ها، گرد کردن مقادیز با دقت معین است فرضا برای مجموعه مقادیر معین {0.93و1.01و3.02و2.99و5.03و5.01و4.98} مقادیر هموارشده برابر خواهدشد .{1.0و1.0و3.0و3.0و5.0و5.0و5.0} = $F_{smoothed}$ البته الگوریتم های هموار کننده همیشه به این سادگی نیستند.

*تفاضل ها و نسبت ها

تغییرات کوچک روی مشخصه ها می توانند بهبود معنی داری را در کارایی کاوش داده ها ایجاد کند .اثرات این تبدیلات کوچک در مشخصه های ورودی/خروجی بویژه در تشخیص و کشف خطا در روش های پیش بینی کننده داده کاوی مهم هستند.

برای مثال ممکن است هدف کنترل های لازم برای فرایند تولید ورسیدن به حالت مطلوب و بهینه باشد.اما بجای نرمال سازی انحراف معیار مشخصه خروجی s(t+1)، شاید با یک حرکت نسبی از مقدار فعلی، مقدار s(t+1)–s(t) در نظر گرفته شود. ویا با بکارگیری s(t+1)/s(t) بعنوان خروجی فرایند داده کاوی موجب بهبود کارایی کل داده کاوی شود.

داده های از دست رفته

در بسیاری از کاربردهای دنیای واقعی کاوش داده ها، حتی با وجود مقدار داده های حجیم و فضای ذخیره سازی مناسب، ممکن است در نمونه های موجود، مقادیری از داده ها از دست

رفتـه(گمشـده) باشـند. در بعضـی روش هـای داده کـاوی ،مقـادیر از دسـت رفتـه داده هـا و فراینـد داده ای مناسب برای رسیدن بـه نتیجـه نهـایی را مـی پذیرنـد (البتـه ایـن نـوع داده کـاوی هـا بیشـتر برای تعداد مقادیر کوچک از دست رفته مصداق دارد)

امـا مشـکل از آنجـا آغـاز میشـود کـه بـرای مجموعـه داده هـای بـزرگ نمـی تـوان از مقـادیر از دسـت رفتـه چشـم پوشـی کـرد. یـک راه حـل بـرای جایگـذینی خودکـار مقـادیر از دسـت رفتـه بـا مقادیر ثابت عبارت است از:

1- جایگزینی تمام مقادیر از دست رفته با یک تک مقدار ثابت سراسری
2- جایگزینی یک مقدار از دست رفته با متوسط مشخصه آن.
3- جـایگزینی یـک مقـدار از دسـت رفتـه بـا متوسـط مشخصـه آن بـرای یـک گـروه مشخص.

نکته ای که بایـد بـه آن توجـه کـرد ایـن اسـت کـه در همـه روش هـای بـالا اشـکال در ایـن اسـت کـه مقدار جایگزین شده مقدار درست و واقعـی آن نیسـت کـه در ایـن حالـت بـه مقـادیر نزدیـک بـه مقـدار از دست داده شده، یک کلاس همگن یا کلاس مصنوعی گفته می شود.

تحلیل داده های نامنطبق

در مجموعه داده های بـزرگ، بـه نمونـه هـایی کـه از رفتـار کلـی مـدل داده ای تبعیـت نمـی کننـد و بطور کلـی متفـاوت یـا ناهماهنـگ بـا مجموعـه باقیانـده داده هـا هسـتند، داده هـای نـامنطبق نامیـده می شوند.

داده هـای نـامنطبق مـی تواننـد توسـط خطـای انـدازه گیـری ایجـاد شـونده یـا نتیجـه نـوع داده ای درونی باشند.

بـرای مثـال اگـر سـن فـردی در پایگـاه داده1- باشـد، مقـدار فـوق قطعـاغلط بـا یـک مقـدار پـیش فـرض فیلـد "سـن ثبـت نشـده "مـی توانـد در برنامـه مشـخص گـردد .ونمونـه هـای دیگـری کـه ممکن است براثر خطای باشند...

و اما راه حل بسیار از الگوریتم های کاوش داده ها سعی در کاهش وبعضا حذف داده های نامنطبق در مراحل پیش پردازش دارند.

نکته در اینجا تحلیلگران داده کاوی باید در الگوریتم های پیش پردازنده بیش از حذف خودکار داده های نامنطبق، بر جنبه های آشکارسازی داده های نامنطبق تاکید کنند.

روش های آماری یکی از ساده ترین راه ها برای تشخیص وکشف داده های نامنطبق روش های آماری است.

بافرض اینکه توزیع مقادیر داده ای انجام شوند، یافتن پارامترهای آماری اساسی مثل مقدار متوسط و انحراف معیار لازم می شوند. براساس همین مقادیر و تعداد داده های نامنطبق، ایجاد مقدار استانه به عنوان یک تابع آماری بوجود می آید.

مثال :مجموعه داده های مشخصه ای مثل سن، با مقدار :20

Age={37و45و55و31و439و28و9و22و41و52و156و39و23و56و3}

میانگین =39.9

انحراف معیار= 45.65

مقدار آستانه یا فاصله =میانگین ± 2* انحراف معیار

همه داده هایی که خارج از بازه [131.2و54.1-]قرار دارند داده های نامنطبق هستند. که با شناخت ویژگی های مشخصه (سن همیشه بزرگتر از صفر) ممکن است که باعث کاهش بیشتر بازه گردد[0و131]در مثال ما مقدار داده های نامنطبق در پایه شاخص معین، 154و139و67- می باشند.

تشخیص داده های نامنطیق برمبنای فاصله روشی دیگر است که محدودیت های ابعادداده ها در روش اماری را ندارد.

روش هـــا و تکنیـــک هـــای برمبنـــای انحـــراف ســـومین کـــلاس از روش هـــای تشـــخیص داده هـــای نامنطبق، دراصل شـبیه سـازی روشـی اسـت کـه مـی تـوان نمونـه هـای غیـر طبیعـی را از یـک مجموعـه نمونه های مشابه دیگر تشخیص داد.

این روش ها ویژگـی هـای اساسـی مجموعـه نمونـه هـا را تعـین و نمونـه هـای متفـاوت از ایـن ویژگـی ها منحرف می شوند و به تبـع آن داده هـای منطبـق آشـکار میشـوند .البتـه بایـد توجـه داشـت کـه تعـین ویژگی هـای مجموعـه نمونـه هـا بسـیار پیچیـده و گـاهی ترکیبـی از انتخـاب طـرح هـای کلـی تـر مـی باشند.

فصل سوم

وب کاوی

۱-۳ وب کاوی

استفاده از وب داده های وب یکی از گام های کلیدی در کشف دانش در پایگاه داده، ایجاد یک مجموعه داده مناسب جهت انجام داده کاوی می باشد.در وب کاوی این داده می تواند از سمت سرور، مشتری، پروکسی سرور یا از یک پایگاه داده سازمان جمع آوری شود.

هر کدام از این داده ها نه تنها از نظ منابع داده متفاوت می باشند بلکه از نظر انواع داده های موجود و محدوده مکانی که آن داده از آنجا جمع آوری می شود و متد پیاده سازی انواع داده ای که در وب کاوی استفاده می شود شامل:

محتوا: داده واقعی در صفحات وب، داده ای که صفحه وب برای نمایش آن به کاربران طراحی شده است.که معمولاً از متن و گرافیک تشکیل شده ولی به آن محدود نمی شود.

ساختار: داده ای که سازمان دهی محتوا را مشخص می سازد. اطلاعات ساختار درون صفحات شامل ترتیب انواع تگ های XML یا HTML در یک صفحه داده شده می باشد و می تواند به صورت یک ساختار درختی نمایش داده شود که تگ ریشه درخت می باشد. اصلی ترین نوع از اطلاعات ساختاری بین صفحات، هایپرلینک[1] است که یک صفحه را به دیگری مرتبط می کند.

استفاده: داده ای که الگوی استفاده از صفحات وب را مشخص می سازد، مثل آدرس هایIP ، رجوع به صفحات و تاریخ و زمان دسترسیپروفایل

کاربر: داده ای که اطلاعات آماری درباره کاربران وب سایت فراهم می سازد که شامل داده ثبت نام و اطلاعات پروفایل مشتری می باشد.

منابع داده داده های استفاده که از منابع مختلفی جمع آوری می شود، الگوهای راهبری از بخش های مختلفی از کل ترافیک وب را نمایش می دهد. جمع آوری در سطح سرورلاگ های وب سرور یک منبع مهم برای اجرای وب کاوی استفاده از وب محسوب می شود زیرا به طور صریح رفتار مرورگری تمام مشاهده کنندگان سایت را ثبت می کند. داده ای که در لاگ سرور ثبت می شود، دسترسی به یک وب سایت که از سوی تمام کاربران صورت می گیرد را منعکس می کند. این فایل های لاگ به فرمت های گوناگونی چون Common log یا Extended log ذخیر می شوند.

[1] hyperlink

جمع آوری در سطح مشتری جمع آوری داده در سطح مشتری می تواند با بکارگیری یک عامل از راه دور (مثل اپلت های جاوا یا جاوا اسکریپت) یا با تغییر کد مرجع یک مرورگر موجود(مثل Mozilla یا Mosaic)پیاده سازی شود.پیاده سازی این نوع روش جمع آوری داده در سطح مشتری به همکاری کاربر در هر دو مورد ذکر شده نیاز دارد.جمع آوری در سطح پروکسییک پروکسی وب به عنوان یک سطح میانی از ذخیره سازی بین مرورگر سمت مشتری و وب سرور محسوب می شود تا زمان بارگذاری صفحه وبی که توسط کاربر تجربه شده را کاهش دهد همانطور که بار ترافیکی در سمت مشتری و سرور را کاهش می دهد.داده های لاگ مربوط به وب معمولاً حجیم و گسترده هستند وبه منظور کشف الگو، این داده ها باید در یک دید یکپارچه، سازگار و جامع جمع آوری شوند . در بیشتر کاربردهای داده کاوی پیش پردازش داده با حذف و فیلتر کردن داده های افزونه و بی ربط و حذف نویزو تبدیل و رفع هر ناسازگاری سروکار دارد. پیش پردازش داده نقش اساسی در کاربردهای کشف دانش در داده استفاده از وب دارا هستند و مهمترین مساله در بیشتر روش های کشف الگو، مشکل آن ها در اداره داده های استفاده از وب در مقیاس بزرگ است . به همین خاطر اکثر فرایندهای KDWUD به طور غیر بر خط انجام می شوند .

تحلیل داده استفاده بدون روش پیش پردازش مناسب نتایج ضعیف و یا حتی خرابی را بدنبال خواهد داشت. بنابراین متودولوژی برای پیش پردازش باید به کار گرفته شود تا هر مجموعه ای از فایل های لاگ وب سرور را به مجموعه ساختاریافته ای از جداول در مدل پایگاه داده رابطه ای تبدیل کند .

فایل های لاگ از روی سایت های مختلف یک سازمان با هم ادغام می شوند تا رفتار کاربرانی که از طریقی ملموس راهبری داشته اند را نمایش دهد . بنابراین این فایل ها باید با حذف درخواست هایی که مورد نیاز نیستند، پاک می شوند مانند درخواست های ضمنی برای آبجکت های تعبیه شده در صفحات وب و یا درخواست هایی که بوسیله مشتری های غیر انسانی وب سایت ها ایجاد می شود . درخواست های باقیمانده با کاربر، نشست های کاربر و مشاهدات و مشاهده صفحات، گروه بندی می شود . و در نهایت مجموعه های پاک و تبدیل شده از درخواست های کاربران در یک مدل پایگاه داده رابطه ای ذخیره می شود .

از فیلتر هایی برای فیلتر کردن داده های بدون استفاده، بی ربط و ناخواسته استفاده می شود تحلیلگر می تواند فایل های لاگ را از وب سرورهای متفاوت جمع آوری کند و تصمیم گیری کند که کدامیک از ورودی ها مطلوب هستند . در واقع هدف این است که اندازه بزرگ داده های استفاده از وب موجود به طور قابل توجهی کاهش یابد و در عین حال کیفیت آن با سازمان دهی آن و فراهم سازی متغیر های یکپارچه اضافی برای تحلیل داده کاوی افزایش یابد .

فرمول بندی مسائل.فرض کنید مجموعه R={ r1+r2+...+rn} مجموعه تمام منابع وباز یک وب سایت باشد . اگر Uمجموعه تمام کاربرانی که به سایت دسترسی داشتند، باشد؛ عنصر لاگ بصورت li تعریف می شود که ui Є U ; ri ЄR است و t زمان دسترسی را نمایش می دهد، s وضعیت درخواست وref iصفحه مورد مراجعه را نمایش می دهد ref i . در برخی ازفرمت های لاگ های وب مثل CLF در حالی که صفحه مورد مراجعه ثبت نشده، اختیاری است s . یک کد سه رقمی است که موفقیت یا شکست درخواست مورد نظر را نشان می دهد . همچنین در موارد دیگر دلیل شکست را نیز بیان می کند . یک وضعیت با مقدار s=200 نشان می دهد که درخواست موفق است در حالی که وضعیت با مقدار s=404 نشان دهنده این است که فایل مورد درخواست در محل مورد نظر یافت نشده است .

li={li1,li2,…,lim}به ترتیب صعودی ذخیره می شوند که یک لاگ وب سرور را تشکیل می دهند . در صورت داشتن N وب سرور، مجموعه فایل های لاگ Log={L1, L2,…,LN} است .با بکارگیری این علائم مسئله پیش پردازش به صورت زیر فرمول بندی می شود .

" با دریافت یک مجموعه از فایل های لاگ مربوط به لاگ های وب سایت های مختلف، کاربر، نشست های کاربر، مشاهده و مشاهدات صفحات کاربران وب سایت در یک بازه زمانی مشخص Δt استخراج می شود ."

پیش پردازش داده

فرایند پیش پردازش گام های زیر را در بر می گیرد :ادغام فایل های لاگ از وب سرورهای گوناگون پاک کردن داده شناسایی کاربران، نشست ها و مشاهده هافرمت بندی داده و خلاصه سازی آن ادغام در ابتدای پیش پردازش داده، درخواست از تمام فایل های لاگ در Log در یک فایل لاگ الحاقی £همراه با نام وب سرور جهت تشخیص بین درخواست های ایجاد شده مربوط به وب سرورهای مختلف وهمچنین توجه به همگام سازی کلاک های وب سرورهای مختلف که از لحاظ زمانی متفاوت اند .

به خاطر دلایل محرمانگی، فایل لاگ نتیجه f را بی نام کرده بطوریکه وقتی که فایل های لاگ به اشتراک گذاشته می شود یا نتایج منتشر می شوند، نام میزبان یا آدرس های IP ، از بین می روند . بنابراین نام اصلی میزبان با یک شناسنده ای که اطلاعاتی درباره محدوده دامنه کد کشور یا نوع سازمان مثل, edu. (.com ,.org) نگهداری می کند، جایگزین می شود .

مسئله ادغام به صورت زیر فرمولبندی می شود:با دریافت یک مجموعه فایل های لاگ Log={L1,L2,…,Ln}این فایل های لاگ در یک فایل لاگ مجزا و منفرد ادغام می شود (فایل لاگ

الحاقی) فرض کنیدLi ، iامین فایل لاگ می باشد Li.c . را به عنوان اشاره گر بر روی درخواست های Liدر نظر بگیرید و Li.1عنصر لاگ جاری از Li است که با Li.c نشان داده می شود وLi.1.time ، زمان tمربوط به عنصرلاگ جاری از Li می باشد و همچنین S=(w1,w2,…,wn) آرایه ای از اسامی وب سرورها می باشد به طوری که S[i] نام وب سرور مربوط به لاگ L i.1 می باشد .

مراحل :

1- مقداردهی اولیه اشاره گر فایل لاگ الحاقی £
2- اسکن عناصر لاگ از هر فایل لاگ Li در Log و افزودن آن به £
3- مرتب سازی عناصر £ به طور صعودی بر اساس زمان دسترسی آن هابرگرداندن مقدار £پاک کردن داده .

گام دوم در پیش پردازش داده حذف درخواست های بدون استفاده از فایل های لاگ می باشد . بطوریکه اگر تمام ورودی های لاگ معتبر نباشند، باید ورودی های بیربط را حذف کنیم .معمولاً این فرایند تمام درخواستهایی که منابع غیر قابل تحلیل مثل تصاویر، فایل های چندرسانه ای و فایل های مربوط به سبک صفحات را در بر می گیرند، را حذف می کند .

برای مثال درخواستهای مربوط به محتوای صفحات گرافیکی تصاویر (*.gif & *.jpg) و همچنین درخواستهای مربوط به هر فایل دیگر در یک صفحه وب یا حتی نشست های راهبری که توسط رباط ها و اسپا یدر های وب انجام می شود .

با فیلتر کردن داده های بی استفاده، می توانیم سایز فایل لاگ را کاهش داده تا از فضای ذخیره سازی کوچکتری استفاده کرده و نیز کارهای بعدی را آسان تر کنیم. برای نمونه، با فیلتر کردن درخواست های تصاویر، سایز فایل های لاگ وب سرور نسبت به سایز اولیه اش تا 50 درصد کاهش می یابد . بنابراین پاک کردن داده حذف ورودی های بی ربطی چون موارد زیر می باشد:

1- درخواستهایی که توسط برنامه های خودکار انجام می شود مثل Web Robot ,Spiders و Crawlerها .این برنامه ها ترافیکی بر روی وب سایت ایجاد می کنند که می توانند بر روی آمار سایت تاثیر بگذارند و همچنین در بررسی هایی که توسط KDWUD انجام می شود مطلوب نیستند .
2- درخواستهای مربوط به فایل های تصویری که به صفحات مشخصی اختصاص داده می شود .

3- درخواست یک کاربر برای مشاهده یک صفحه خاص معمولاً در چندین در چندین عنصر از لاگ منعکس می شود زیرا هر صفحه گرافیک هایی را شامل می شود که فقط آنهایی برای ما مهم هستند که کاربر صریحاً آنها را درخواست کرده که معمولاً فایل های منتی هستند .

عناصر با کدهای وضعیت HTTP نا موفق . کدهای وضعیت HTTP برای نشان دادن موفقیت یا شکست یک درخواست بکار می روند که در اینجا ما فقط عناصر با کد بین 200 تا 299 که با موفقیت انجام شده اند در نظر می گیریم .

عناصری که متدی به غیر از GET و POST دارند .شناسایی در این گام درخواستهای غیر ساختیافته یک فایل لاگ به صورت کاربر (user)، نشست کاربر (user session)، مشاهدات و ملافات صفحات (page view, visit)گروه بندی می شود . در پایان این گام فایل لاگ به صورت یک مجموعه از تراکنش ها خواهد بود (نشست کاربر یا مشاهدات)کاربردر بیشتر موارد فایل لاگ فقط آدرس های کامپیوتر نام یا IPو عامل کاربر را فراهم می سازد به عنوان مثال فایل های لاگ . ECLF برای وب سایتهایی که نیازمند ثبت کاربرهستند، فایل لاگ همچنین User login را شامل می شود (به عنوان سومین رکورد در یک عنصر لاگ) که برای شناسایی کاربر استفاده می شود . وقتی که user login موجود نباشد هر IP به عنوان کابر در نظر گرفته می شود . با این حال این واقعیت وجود دارد که یک آدرس IP توسط چندین کاربر استفاده می شود واین برای KDWUD جهت شناسایی کاربر کافی نیست . به هر حال هنوز هم مکانیزمی برای تشخیص و تمایز بین کاربران برای تحلیل رفتار دسترسی کاربر مورد نیاز است .

نشست کاربر شناسایی نشست کاربر از فایل لاگ بدلیل پروکسی سرورها، آدرس های پویا و مواردی که چندین کاربر از طریق یک کامپیوتر دسترسی پیدا می کنند (در کتابخانه، کافی نت و...) یا یک کابر از چندین مرورگر یا کامپیوتر استفاده می کند، امکان پذیر نمی باشد . یک نشست کاربر به صورت ترتیبی از درخواست ها که بوسیله یک کاربرمنفرد در یک دوره زمانی مشخص تعریف می شود . یک کاربر می تواند یک (یا چند) نشست در طول یک دوره زمانی داشته باشد .شناسایی نشست عبارت است از فرایند قطعه بندی لاگ دسترسی هر کاربر به نشست های دسترسی مجزا .دو روش بر اساس زمان وجود دارد که شامل روش مبتنی بر طول نشست (Session-duration) و روش مبتنی بر page-stay-time همچنین می توانیم از یک آستانه زمانی timeout استفاده می کنیم .

۲-۳ روش‌های وب‌کاوی

داده‌های موجود در وب، انواع مختلفی دارند. این انواع داده‌ها را می‌توان به سه گروه کلی تقسیم کرد:

1- محتوای وبWeb Content ، مانند متن، تصاویر، جداول و ... .
2- کاربرد وبWeb Usage ، مانند فایل‌های LOG بدست آمده در سرورها و ... .
3- ساختار وبWeb Structure ، مانند لینک‌ها و برچسب‌ها.

حال با توجه به تقسیم‌بندی بالا، وب‌کاوی به سه دسته کلی ذیل تقسیم می‌گردد:

1- وب‌کاوی محتواWeb Content Mining
2- وب‌کاوی کاربردWeb Usage Mining
3- وب‌کاوی ساختارWeb Structure Mining

در سه بخش پیشِ رو به بررسی مختصر هر یک از این دسته‌ها خواهیم پرداخت .

Web Content Mining وب‌کاوی محتوا، فرآیند استخراج اطلاعات کاربردی از محتوای وب را گویند. محتوای وب، داده‌های ساخت‌یافته (Structured) مانند جدول، نیمه ساخت‌یافته مانند برچسب‌های HTMLو غیرساخت‌یافته (Unstructured) مانند متن معمولی را شامل می‌شود.این حیطه از روش‌های دیگری مانند داده‌کاوی متن(Text Mining) ، پردازش زبان طبیعی (NLP) و بازیابی اطلاعات (Information Retrieval)در دل خود استفاده می‌نماید. در وب‌کاوی محتوا تکنیک‌های داده‌کاوی معمولی مثلClassification ، Clusteringو Association را نیز می‌توان به کار گرفت .

Web Usage Mining استخراج الگوهای معنادار از داده تولید شده در سرورها، پست الکترونیکی و ... را وب‌کاوی کاربرد گویند. با کمک این روش می‌توان رفتار کاربران مختلف را پیش‌بینی کرده و بر اساس آن، صفحات وب را برای آن کاربر خاص، شخصی‌سازی نمود.با استفاده از وب‌کاوی کاربرد می‌توان به شخصی‌سازی صفحات وب پرداخت(Personalization) ، سایت‌ها را اصلاح نمود(Site Modification) و یا کارائی سایت‌ها افزایش داد

نکته دیگر در باره فرمت فایل‌های LOG است. این فایل‌ها فرمت استاندارد خاصی ندارند و در هر کاربرد و هر سرور بر حسب نیاز خود، فرمت و قالب مخصوصی را تعریف کرده و از آن قالب خاص استفاده می کنند.

Web Structure Mining وب‌کاوی ساختار به استخراج اطلاعات از ساختار صفحات وب می‌پردازد. این روش خود به دو دسته‌ی Intra-page و Inter-page تقسیم می‌شود.

در مورد روشIntra-page ، ساختار درونی یک صفحه وب، در نظر گرفته می‌شود، در صورتی که در روش Inter-pageساختار صفحات وب در ارتباط با یکدیگر مورد توجه است. به روشInter-page ، تحلیل ابرلینک‌ها (Hyperlink Analysis) نیز گفته می‌شود.در این روش معمولا صفحات وب و لینک‌های میان آنها به شکل گراف جهت‌دار نمایش داده می‌شود. هر گره از این گراف، نشان دهنده یک صفحه وب بوده و هر یال جهت‌دار از گره "الف" به گره "ب" وجود لینکی از صفحه "الف" به صفحه "ب" را نمایش می دهد.

فصل چهارم

مدیریت ارتباط با مشتری

۱-٤ تاریخچه:

شاید بتوان تاریخچه ظهور مباحث مرتبط به CRM را در سه دوره زیر خلاصه نمود:

الف) دوره انقلاب صنعتی (تولید دستی تا تولید انبوه) ابتکارات فورد در بکارگیری روش تولید انبوه به جای روش تولید دستی، یکی از مهمترین شاخص‌های این دوره می باشد. هر چند تغییر شیوه تولید باعث شد که محدوده انتخاب مشتریان از نظر مشخصه‌های محصول کاهش یابد (نسبت به تولیدات صنایع دستی) اما محصولات تولید شده به روش جدید از قیمت تمام شده پایین تری برخوردار شدند. به عبارتی دیگر در انتخاب روش تولید انبوه از سوی فورد، افزایش کارایی و صرفه اقتصادی مهمترین اهداف پیش بینی شده بودند.

ب) دوره انقلاب کیفیت (تولید انبوه تا بهبود مستمر) این دوره هم‌زمان با ابتکار شرکت‌های ژاپنی مبنی بر بهبود مستمر فزآیندها آغاز شد. این امر به نوبه خود به تولید کم هزینه تر و با کیفیت تر محصولات منجر شد. با مطرح شدن روشهایی نوین مدیریت کیفیت مانند TQM این دوره به اوج خود رسید. اما با افزایش تعداد شرکت‌های حاضر در عرصه رقابتی و گسترش فرهنگ حفظ و بهبود کیفیت محصول (از طریق ابزارهای مختلف کیفیتی)دیگر این مزیت رقابتی برای شرکتها پیشرو و کارساز نبوده و لزوم یافتن راه‌های جدیدی برای حفظ مزیت رقابتی احساس می شد.

ج) دوره انقلاب مشتری (بهبود مستمر تا سفارشی‌سازی انبوه) در این دوره با توجه به افزایش توقع مشتریان، تولید کنندگان ملزم شدند محصولات خود را با هزینه کم، کیفیت بالا و تنوع زیاد تولید کنند. به معنای دیگر تولید کنندگان مجبور بودند توجه خود را از تولید صرف به یافتن راه‌هایی برای رضایت مشتریان سابق خود معطوف نمایند.

٢-٤ مشتری کیست

کسی که محصولات یا خدمات سازمان را خریده و یا از آن استفاده میکند . به عنوان مشتری تعریف میگردد . در مجموعه میتوان مشتریان را به دو دتسه تقسیم کرد . مشتریان خارجی و مشتریان داخلی

مشتریان خارجی در بیرون از سازمان بوده و محصولات و خدمات میخرند . تمام کارمندانی که به نوعی در فر آیند های تهیه و توضیح کالا یا خدمات نقش دارند باید اثر کار و وظیفه خود را در بالابردن سطح رضایت مشریان خارجی به خوبی درک نمایند .

در برابر مشتریان خارجی هر سازمانی مشتریان داخلی دارند که به اندازه مشتریان خارجی مهم اند . در تمام مراحل از قسمت های مهندسی گرفته تا تولید و دیگر فر آیند ها همواره یه مشتری داخلی و جود دارد که محصول و خدمات را دریافت میکنند و در عوض محصول خدمتی را ارایه مینماید

٣-٤ مدیریت ارتباط با مشتری

مدیریت ارتباط با مشتری یک فرایند تجاری است که تمام جوانب مشخصه های مشتری را آدرس دهی می کند، دانش مشتری را به وجود می آورد، روابط را با مشتری شکل می دهد و برداشت آنها را از محصولات یا خدمات سازمان ایجاد می کند. مدیریت ارتباط با مشتری توسط چهار عنصر از یک چارچوب ساده تعریف شده است: دانش، هدف، فروش و خدمت.

مدیریت ارتباط با مشتری با درنظر گرفتن اینکه چه محصولات یا خدماتی، به چه مشتریانی، در چه زمانی و از طریق چه کانالی عرضه شود، بهبود را در پی خواهد داشت. این مدیریت از اجزای مختلفی تشکیل شده است.

پیش از اینکه فرایند آن آغاز شود، شرکت باید اطلاعات مشتری را در اختیار داشته باشد. این اطلاعات می تواند از داده های داخلی مشتریان و یا از داده های منابع خارجی خریداری شده، به دست آید. برای داده های داخلی منابع مختلفی وجود دارد مانند پرسشنامه ها و بلاگ ها ، سوابق کارت اعتباری و....

منـابع داده خـارجی یـا بانکهـای داده خریـداری شـده ماننـد آدرسـها، شـماره تلفـن هـا، پروفایـل هـای بازدید از وب سایتها کلیدی برای به دست آوردن دانش بیشتری از مشتری است.

بیشـتر شـرکتها، بانکهـای داده ای عظیمـی شـامل داده هـای بازاریـابی، منـابع انسـانی و مـالی را دارا هسـتند. بنـابراین، سـرمایه گـذاری در زمینـه انبـار داده، یکـی از اجـزای حیـاتی در اسـتراتژی مـدیریت ارتباط با مشتری است.

پس از تهیه و تخصیص منابع داده، سیستم مدیریت ارتباط با مشتری باید با به کارگیری ابزارهایی مانند داده کاوی، داده ها را تجزیه و تحلیل کند. اعم از اینکه شرکت تکنیک های آماری سنتی را به کار می برد یا یکی از ابزارهای نرم افزاری مانند داده کاوی را، کارشناسان نیاز به فهم داده های مشتری و روابط تجاری دارند. بنابراین، داشتـــن افرادی متخصص که این داده ها را با ابزارهای مربوطه استخراج و به صورت اطلاعات درآورند، مهم است

٤-٣ انواع CRM:

4-3-1 CRM عملیاتی (Operational):

در ایـن روش کلیـه مراحـل ارتبـاط بـا مشـتری، از مرحلـه بازاریـابی و فـروش تـا خـدمات پـس از فروش و اخـذ بـازخورد از مشـتری، بـه یـک فـرد سـپرده مـی شـود، البتـه بـه نحـوی کـه فروشـندگان و مهندسان ارائه خـدمات بتواننـد سـابقه هـر یـک از مشـتریان را بـدون مراجعـه بـه ایـن فـرد در دسـترس داشـته باشـند.از ابـزار و روشـهای CRM عملیـاتی میتـوان بـه SFA یـا قـدرت فـروش مکـانیزه اشـاره نمود که کلیـه عملیـات بـه مـدیریت تمـاس بـورس و مـدیریت اداره فـروش را برعهـده دار CSS ابـزار دیگـر CRM عملیـاتی اسـت کـه در آن بـه جـای ارتبـاط تلفنـی بـا مشـتری، از ابزارهـای دیگـری ماننـد ارتبـاط رو در رو، اینترنـت، فـاکس و کیوسـک‌هـای مخصـوص پاسـخگویی بـه مشـتریان اسـتفاده مـی شود.

4-3-2 CRM تحلیلی (Analytical):

در CRM تحلیلـی ابزارهـا و روشـهایی بـه کـار مـی رود کـه اطلاعـات بـه دسـت آمـده از CRM عملیاتی را تجزیـه و تحلیـل نمـوده و نتـایج آن را بـرای مـدیریت عملکـرد تجـاری آمـاده مـی کنـد. ایـن سیسـتم مهمتـرین نـوع از CRM مـی باشـد. بـه ایـن صـورت کـه شـامل داده‌هـایی اسـت کـه برنامـه‌هـا

جهت بر قراری ارتباط با مشتری به آن نیاز دارند. به عبارت دیگر این داده‌های خام در اختیار برنامه‌های CRM قرار می گیرند و پس از کار بر روی این داده‌ها، نتیجه مناسب در اختیار شرکت و مشتری قرار داده می شود. اما اگر بخواهیم یک تعریف کامل ارائه نماییم: بدست آوردن، ذخیره، پردازش، تفسیر و ارائه گزارش به استفاده کنندگان داده‌ها مشتری می باشد. شرکتهای زیادی هستند که این داده‌ها را جمع آوری کرده و پس از استفاده از الگوریتمهایی مختلف سعی در تحلیل و تفسیر این داده‌ها می نمایند.در واقع، CRM عملیاتی و تحلیلی در یک تعامل دو طرفه هستند.

CRM 3-3-4 تعاملی (Collaborative):

در این نوع ارتباط، مشتری برای برقراری ارتباط با سازمان از سهل‌ترین روش ممکن مانند تلفن، تلفن همراه، فاکس، اینترنت و سایر روش‌های مورد نظر خود استفاده می نمایند. این نرم‌افزارها را PRM می نامند. CRM تعاملی به دلیل امکان انتخاب روش از سوی مشتری و اینکه اکثر فرآیندها (از جمع آوری داده‌ها تا پردازش و ارجاع مشتری) در حداقل زمان ممکن به مسوول مربوطه صورت می گیرد باعث مراجعه مجدد مشتری و ادامه ارتباط با شرکت می‌شود

٤-٤ نقش مدیران ارشد در ارتباط با مشتریان

یک شرکت هواپیمایی را در نظر بگیرید. اگر این شرکت به لحاظ ارتباط با مشتریان ضعیف باشد، به آسانی می‌توان آن را زیر سوال برد و از آن انتقاد کرد.

جان تگ، مدیر ارشد عملیاتی شرکت هواپیمایی یونایتد ایرلاینز قسمت اعظم کار خود را روی ارتقای سیستم‌های اطلاع‌رسانی به مشتری، تکنولوژی‌های مشتری‌مدار و قابلیت‌های خدمات به مشتری متمرکز کرده است. وی در مصاحبه‌ای بیان داشت» بزرگ‌ترین چالش ما آگاه کردن سازمان به میزان فعالیت لازم برای ارتباط بهتر با مشتری است. بدون احساس ضرورت، حقیقتا تغییر دادن فرهنگ مان بسیار سخت است. حال که یونایتدایرلاینز با شرکت هواپیمایی Continental ادغام شده است امید آن می‌رود که قدرت ارتباطی آن بیشتر شود .دلیل دوم که چرا مدیران شرکت‌های بزرگ ارتباط محدودی بامشتریان دارند، این است که آنها قدرت نمادین روش‌های مدیریتی خود را دست کم یا نادیده می‌گیرند. نمونه و الگو بودن جزو ماهیت ثانویه مدیران کارفرما است و مدیران ارشد شرکت‌های بزرگ باید نسبت به این جنبه از مدیریت آگاهی بیشتری داشته باشند .مدیران ارشد نمی‌توانند وظیفه فرهنگ‌سازی در سازمان را به فرد

دیگری واگذار کنند. آنها موثرترین و مهم‌ترین شخص در شرکت‌ها هستند که قادر به انجام این کار هستند. شیوه‌های مدیریتی مدیران ارشد روی عملکرد تیم‌های کاری، کارمندان و مدیران دیگر تاثیر می‌گذارد. زمانی که ارتباط مدیران ارشد با مشتریان قطع شود برای کل سازمان نیز چنین اتفاقی رخ می‌دهد .سومین دلیلی که ارتباط با مشتری به مساله مبهمی‌تبدیل می‌شود این است که به عنوان یک مساله مهم مدیریتی در نظر گرفته نمی‌شود. امروزه دغدغه و نگرانی مدیران مدرن مسائلی چون استراتژی، ادغام و تملک، قیمت سهام شرکت و قدرت ترازنامه است. ارتباط با مشتری کار سطح پایینی تلقی می‌شود و آنقدر پیش پا افتاده در نظر گرفته می‌شود که نمی‌توان آن را جزء اصلی فرمول موفقیت سازمان در نظر گرفت .چهارمین و آخرین دلیلی که شرکت‌های بسیاری وجود دارند که ارتباط مناسبی با مشتریان ندارند این است که اکثر مدیران از تمام جنبه‌ها به مشتریان فکر نمی‌کنند. از دید یک مدیر ارشد در شرکتی بزرگ، مشتریان معمولا اطلاعات، نمودار یا اعداد و ارقام هستند. آنها به عنوان انسان‌های واقعی دیده نمی‌شوند .نقطه کور مدیریت کوتاهی در ایجاد سازمان‌های مشتری‌مدار و عدم درک نقشی که مدیریت ایفا می‌کند اشتباه بزرگ و هزینه‌برداری است. عدم ارتباط حقیقی و پایدار منجر به نابودی عواملی چون نوآوری، وفاداری مشتری و رشد سودآور در شرکت‌ها شده است. مدیران اندکی دریافته‌اند که چگونه سازمان‌های مشتری مدار را برپا کنند و فقدان چنین مدیریتی میلیاردها دلار سرمایه را در اقتصادهای جهانی و آمریکا از بین برده است. بارزترین مثال، صنعت خودروسازی ایالات‌متحده است که به دلیل عدم ارتباط مناسب مدیران با مشتریان فلج شد .عدم برقراری ارتباط با مشتریبرای بسیاری از صنایع دیگر نیز تبدیل به مساله بزرگی شده است. شرکت‌های کرایه اتومبیل در حال تلاش برای متعادل کردن کاهش هزینه‌ها و خدمات به مشتری هستند. این وضعیت در مورد بیمارستان‌ها، شرکت‌های هواپیمایی و بانک‌ها نیز همین‌طور است. تعداد اندکی در این صنایع دارای مدیران مشتری مدار هستند و آنهایی که ارتباط مناسبی با مشتریان دارند (برای مثال شرکت هواپیمایی Southwest به داشتن مشتریان وفادار، تقویت نوآوری و لذت بردن از کسب و کار پایدار مشهور هستند) .نادیده گرفتن اهمیت ارتباط با مشتریپیامد قابل توجه دیگری نیز داشته است. سال 2008 در زمان رکود اقتصادی شرکت‌های آمریکایی بیش از 130 میلیارد دلار برای آموزش و رشد کارمندان هزینه کردند که از این میان، حدود 50 میلیارد دلار صرف ارتقا و پیشرفت مدیران شد. این میزان مبلغی است که شرکت‌ها

هزینه کردند تا از داشتن مدیران مستعد و مجرب برای آینده‌ای رو به رشد اطمینان حاصل کنند. یکی از رایج‌ترین و موثرترین راه‌ها برای پرورش مدیران آینده سنجیدن مهارت‌های مدیریتی آنها و ارائه بازخورد به مدیران است؛ یعنی اینکه به آنها نشان دهد تا چه میزان این مهارت‌ها را دارا هستند. به منظور انجام این کار، شرکت‌ها باید دقیقا تصمیم بگیرند چه مهارت‌هایی مهم‌ترهستند و ارزش ارزیابی دارند. این لیست از مهارت‌ها اغلب به عنوان «برنامه موفقیت» تلقی می‌شوند.راه‌حل چیست؟ در اینجا به ذکر چند راهکار برای ایجاد ارتباط مناسب با مشتریان می‌پردازیم. جای هیچ تعجبی نیست که این راهکارها همان مشخصه‌های مدیران موفق است :به ایجاد تنظیمات سازمانی جدید بپردازید. در شرکت‌های کوچک، تشکیلات سازمانی ساده است و برای مدیران ارتباط نزدیک با مشتری آسان است. در شرکت‌های بزرگ‌تر، تشکیلات سازمانی دارای پیچیدگی است. نظام سلسله‌مراتب و کاغذ بازی مانع رسیدن صدای مشتری به مدیران می‌شود. مدیران ارشد به دلیل نداشتن دید روشن قادر به تشخیص یا پاسخ سریع به مشتریان نیستند .در شرکت‌های متوسط، یک راه موثر برای از بین بردن موانع سازمانی، حذف دیوارها به طور فیزیکی بین سطوح است. شرکت Zappos با بیش از دوهزار کارمند دست به چنین اقدامی‌زد: در دفتر مرکزی پارتیشن‌هایی که میزکار کارکنان را از قسمت‌های دیگر مجزا می‌کند به حدی کوتاه است که افراد می‌توانند بایستند و به راحتی با آنها صحبت کنند. شرکت Communispace نیز یکی از سازمان‌هایی است که به ارتباط با مشتری بسیار اهمیت می‌دهد. دفتر مرکزی این شرکت در ایالت ماساچوست آمریکا دارای طراحی فضای باز است. خانم دیان حسان، مدیر ارشد این شرکت طی مصاحبه‌ای گفت: نمی‌توانم شکل دیگری از محیط کار را تصور کنم. این مدل با فرهنگ، ارزش‌ها و تعهد ما به شفافیت و قابل دسترس بودن مطابقت دارد.اقدامات مشتری مداری را ارزیابی کنید و برای آن پاداش قائل شوید. سیستم‌های ارزیابی و پاداش، چه رسمی‌و چه غیررسمی کمک می‌کنند تا تعیین کنید افراد به چه چیزی توجه می‌کنند و چگونه رفتار می‌کنند. مدیران ارشد باید در دادن پاداش به افرادی که برای ارتباط با مشتری تلاش بسیار می‌کنند پیشقدم باشند. یک گفته مشهور می‌گوید «هرآنچه که ارزیابی شود، مدیریت می‌شود» و این دستورالعمل در هر جنبه‌ای از ارتباط با مشتری کاربرد دارد. مساله اصلی این است که چه چیز باید مورد سنجش و ارزیابی قرار بگیرد و این ارزیابی چگونه باید انجام شود. در صورت امکان، معیارهای سنجش باید فراتر از داده‌های کمی و آماری سنتی

حاصل از تحقیقات بازار باشد. مدیران ارشد باید اصرار داشته باشند که صدای مشتری شنیده شود و کوچک‌ترین جزئیات آن شرح داده شود. داده‌های کمی و آمار و ارقام مفید هستند اما معیارهای اندازه گیری کیفی و به صورت حکایت و داستان دارای عمق بیشتری است. صحبت و گفت‌وگو بیشتر از اعداد و ارقام به ما اطلاعات می‌دهند.از تکنولوژی آنلاین استفاده کنید. با توجه به اندازه، تنوع و ماهیت جوامع مشتریان، ارتباط با آنها به شیوه‌ای اصولی و متفکرانه بدون استفاده از تکنولوژی آنلاین تقریبا غیرممکن است. اینترنت این امکان را می‌دهد که سازمان‌ها، اطلاعات بسیار زیاد و همین طور ارزشمندی را جمع‌آوری کنند. مدیران ارشد شرکت‌ها به جای عقب کشیدن و قطع ارتباط با میلیون‌ها مشتری باید از فناوری اینترنت استفاده کنند و اصرارداشته باشند که سازمان‌ها برای بهترشنیده شدن صدای مشتری از این تکنولوژی بهره ببرند. همچنین تکنولوژی آنلاین برای مدیران پرمشغله این امکان را فراهم می‌کند تا بر موانع زمانی غلبه کنند .ضرورت فرهنگی سازمان‌هایی که بامشتریان خود ارتباط دارند، جو سازمانی به‌خصوصی دارند. این موضوع اصلا ربطی به اندازه یا نوع فعالیت شرکت ندارد، چنین محیط‌های کاری دارای ویژگی‌های مهم مشترکی هستند: آنها خیلی ساختارگرا نیستند؛ دارای استاندارد بالا هستند؛ برای مسوولیت فردی اهمیت بسیاری قائلند؛ روحیه همکاری بالایی دارند؛ دارای «سیستم باز» هستند یعنی با محیط اطراف خود تعامل دارند و در آخر تعهد بسیار بالایی به شرکت، اهداف آن و کارمندانش دارند .مدیران چگونه باید این جو سازمانی را ایجاد کنند؟ برای شکل‌گیری فرهنگ ارتباط با مشتریچه اقدامات مدیریتی مورد نیاز است؟ فرهنگ‌سازی صحیح وابسته به برخی اقدامات مدیریتی است که با محیط سازمانی دارای عملکرد بالا و موفقیت کسب و کار همراه شده است. به چند مورد از این اقدامات اشاره‌ای خواهیم کرد:

1- اهداف چالش انگیز را برای ارتباط با مشتریان قرار دهید. این به معنی وادار کردن سازمان‌ها به اتخاذ اهداف بزرگ و به چالش کشیدن سازمان برای بهبود ارتباط با مشتری است
2- تعهد شخصی خود را برای دستیابی به اهداف ایجاد ارتباط با مشتری نشان دهید. این بدان معنا است که مدیران باید کاری را که به دیگران توصیه می‌کنند ابتدا خود آن را انجام دهند

3- به هنگام برقراری ارتباط با مشتری، از نوآوری و انجام ریسک‌های حساب شده استقبال کنید. زمانی که کارکنان سازمان برای انجام کارهای نو و بدیع در پاسخ بهمشتریان احساس آزادی کنند، فرهنگ ارتباط با مشتری بیشتر تقویت می‌شود

4- برای حفظ ارتباط بامشتری مسوولیت‌های هر فرد را روشن کنید. مدیران باید راهنمایی و هدایت کافی برای هماهنگ سازی فعالیت‌های مختلف را فراهم کنند و مطمئن شوند که اطلاعات مشتری به افراد مناسبی منتقل می‌شود

5- افراد را به مشارکت در تصمیم‌گیری‌های مربوط به ارتباط با مشتریان تشویق کنید. مدیران باید به تمام کسانی که به اطلاعات مشتریان دسترسی دارند، اختیار و قدرت تصمیم‌گیری بدهد .علاوه بر این اقدامات، مدیرانی که در ایجاد فرهنگ ارتباط با مشتری موفق عمل کردند گرایشات و جهت‌گیری‌های دیگری نیز داشته اند. آنها در کنار توجه به مشتری روی رقبا، عوامل موثر در بازار و جامعه متمرکز شده اند. سعی آنها بر این است که از بوروکراسی، سیاست‌های اداری و هر آنچه مخل عملکرد سازمان می‌شود اجتناب ورزند.

عامل اصلی در شکل‌گیری فرهنگ ایجاد ارتباط با مشتری اندازه یا پیچیدگی سازمان‌ها نیست بلکه خواست و مهارت مدیران ارشد نقش اصلی را ایفا می‌کند

٤-٥ مزایای بکارگیری سیستم‌های مدیریت ارتباط با مشتری

CRM یک واقعیت ملموس برای سازمان‌های تجاری است و به طور خلاصه مزایای زیر را برای سازمان به دنبال دارد:

١ - پاسخگویی سریع به درخواست مشتریان
٢ - فراهم کردن شرایط مراجعه مجدد مشتری
٣ - کاهش هزینه‌های تبلیغاتی
٤ - افزایش فرصت‌های بازاریابی و فروش
٥ - شناخت عمیق تر مشتری
٦ - دریافت باز خورد از مشتری و توسعه خدمات و محصولات جاری

٦-٤ چارچوب گارتنر

طرح‌های CRM نیازمند چارچوبی اند که تضمین کند برنامه‌های سازمان در مبنای استراتژیک ویکپارچه در نظر گرفته می شوند. گارتنر چنین رویکردی را که شامل 8 گام است طراحی کرده است:

1- تدوین چشم انداز سازمان ایجاد چشم اتدازموثر مستلزم این است که رهبران سازمانی: • معانی CRMرا برای موسسه تعریف کنند • اهداف را تعیین کنند • تصویری از آنچه سازمان می خواهد برای مشتریان هدفش باشد ترسیم کنند

هدف: باید مجموعه‌ای از ارزش‌های متمایز شده که برای مشتریان مهم است خلق گردد.

2- تدوین استراتژیهای CRM استراتژی CRM نگرشی را در مورد نحوه ایجاد ارتباط با مشتریان ارزشمند و نحوه وفاداری در آنان فراهم می آورد. گام اول: تدوین استراتژی CRM بخش بندی مشتریان در گروه‌ها، تعیین اهداف و معیارهای سنجش برای هر بخش است. گام دوم: ارزیابی وضعیت پایگاه مشتری به عنوان یک دارایی است. این کار از طریق ترسیم نمودار قوت‌ها و ارزش ارتباط با مشتریان در دو بعد صورت میگیرد:

• مشتری تا چه اندازه برای سازمان ارزشمند است؟

• سازمان تا چه اندازه برای مشتری ارزشمند است؟

3- طراحی تجربه مشتری در این مرحله باید اطمینان حاصل شود که محصولات و تعاملات سازمان باعث خلق ارزش برای مشتریان گشته، به طور پایدار ارائه شده و به موقعیت به بازار مطلوب دست پیدا کرده است یا خیر اجرای سیستم باز خورد عملیاتی سبب افزایش آگاهی سازمان از شکایات مشتری گشته، حل این شکایات را میسرمی سازد.

4- میسر ساختن همکاری سازمانی همکاری سازمانی به معنی تغییر فرهنگ، ساختارهای سازمانی و رفتارهاست تا اطمینان حاصل شود که کارکنان، شرکا و تامین کنندگان در جهت ایجاد ارزش برای مشتریان با یکدیگر همکاری می کنند.

5- طراحی مجدد فرآیندهای کسب و کار در زمینه طراحی مجدد فرآیندهای مربوط به مشتری استفاده از چارچوب زیر الزامی است: • نقاط تماس و فرآیندهایی را که بر روی مشتریان تاثیر می گذارند، حسابرسی کرده، نقشه آنها را ترسیم کنید. • فرآیندهای کلیدی را از دیدگاه مشتری شناسایی کنید و فرآیندهایی که می توانند بیشترین نارضایتی را ایجاد کنند پیدا کرده، در وحله اول بر آنها تمرکز شود.

• این فرآیندها را بر اساس اثراتشان بر روی هدف CRM مشخص و اولویت بندی کنید.

• تغییرات لازم را در سازمان اعمال کنید (هیچ فرآیندی را نباید بدون مجری و مسئول رها کرد)

• با استفاده از اهداف مشتریان، اهداف قابل اندازه گیری و یا مفهومی را تعیین کنید. برای هر فرآیند کلیدی نوعی توافقنامه سطح خدمت به مشتری را برقرار کنید.

• فرآیندها را بر اساس اهمیت آنها به مشتریان و اثرات آنها بر اهداف CRM الویت بندی کنید.

6- تدوین استراتژی اطلاعات مشتری منظور از این مرحله جمع آوری داده‌های صحیح و ارسال آنها به مکان صحیح است. مدیریت موفق ارتباط با مشتری نیازمند خلق نوعی ((عرضه خون)) اطلاعاتی است که در سراسر سازمان جریان یافته و سیستم‌های عملیاتی وتحلیلی را یکپارچه کند.

7- استفاده از فناوری منظور از این مرحله مدیریت داده‌ها و اطلاعات، برنا مه‌های کاربردی پیش روی مشتری، زیر ساخت‌ها و معماری IT است.

8- معیارهای سنجش منظور از این مرحله اندازه گیری شاخص‌های درونی و بیرونی موفقیت و شکست CRM است.

این شاخص‌ها دارای کاربرد زیر هستند:

• تعین و اندازه گیری سطح تحقق اهداف CRM

• ارائه باز خورد برای اصلاح استراتژی CRM واجرای آن

• نظارت بر تجربه مشتری از سازمان

• تغیر شیوه جبران خدمات کارکنان و مشوق‌های داده شده

• ارزیابی سازمان نسبت به رقبا

در گام آخر: اهدافی را که باید برآورده شوند و تاکتیک هایی را که باید مورد استفاده قرار گیرند تعریف می کنیم.

٤-٧ اندازه گیری رضایت مشتری(CSM)چیست؟

اندازه گیری رضایت مشتریان یکی از فعالیت های بسیار مهم برای هر سازمان است. سازمانی که بتواند رضایت مشتریان خودش را اندازه گیری نماید، در حقیقت مانند این است که علائم حیاتی خودش را دائماً پیش رو دارد، و می تواند آنها را کنترل کند و با توجه به آنها سازمان خودش را به پیش ببرد.

مروزه در دنیای مدرن و مدیریت مدرن اندازه گیری رضایت مشتریان از پیشرفت الگوی مدرن تبدیل به یک ضرورت شده است

اهمیت CSMدر استاندارد بند 1-2-8- استاندارد ISO 9001 بیان می دارد :

به عنوان یکی از طرق اندازه گیری عملکرد سیستم مدیریت کیفیت، سازمان بایست اطلاعات مرتبط با برداشت و درک مشتریان را پالایش نموده و تعیین نمایند که آیا سازمان توانسته نیازمندی‌های مشتریان را برآورده سازد.ضمنا" متدهای دستیابی و به کارگیری این اطلاعات نیز بایستی تعیین گردند .

4-7-1 نحوه اندازه گیری رضایت مشتری چگونه است؟

1- نمونه گیری: (در نظر گرفتن تمام جوانب) پراکندگی جغرافیایی و پراکندگی از لحاظ بازار
2- ابزار تحقیق: تعیین شاخص‌های رضایت که مهم‌ترین عناصر در رضایت مشتریان را اندازه گیری کند .
3- متدولوژِی تحقیق: انتخاب روش آماری و تحلیل با هدف آسان شدن عددی کردن شاخص‌های کیفی
4- تشخیص منابع: تعریف منابع و استراتژِی سازمان پس از تحلیل در فعالیتهای شرکت

اثرات رضایت مشتری چیست؟

رضایت مشتری از سه طریق به افزایش درآمد و سود منجر می شود. تکرار خرید مشتری، خرید کالای جدید و خرید کالا توسط مشتریان جدیدی که توسط مشتریان راضی به کالا تمایل پیدا کرده اند. امروزه در کشورهای صنعتی برنامه های ارتباط با مشتریان در سرلوحه برنامه های

بازاریـابی عرضـه کننـدگان قـرار گرفتـه اسـت. امـروزه دیگـر هـیچ تولیدکننـده و عرضـه کننـده ای به فروش یک بار به مشتری نمی اندیشد .

نکته مهم این است که در صورت نارضایتی مشتری، تمامی مکانیسم ها در جهت عکس عمل کرده و درآمد و سود عرضه کننده را کاهش می دهد. بررسیها نشان می دهد که مشتریان ناراضی در انتقال احساس خود به دیگران فعال تر و موفق تر عمل می کنند .

٤-٨ تفاوت CRM وCSM چیست؟

هـدف یـک سیسـتم مـدیریت روابـط بـا (مشـتریان ،داشـتن مشـتریان وفـادار اسـت). یعنـی شـما بـا اجرای CRM در سازمانتان انتظـار داریـد مشـتریانی راضـی تـر (حتـی اگـر کمتـر از قبـل هسـتند) داشـته باشید تا پول بیشتری ازمشتریان کمتـر بدسـت بیاریـد. سـپس حتمـا مـی خواهیـد تعـداد آنهـا را افـزایش بدهید. این در درازمدت به شما کمک میکند که مشتریانی راضی و بسیار داشته باشید .

ولـی CSM در ارتبـاط بـا میـزان رضـایتمندی مشـتریان شـما، صـرفا یکـی از عوامـل (البتـه از عوامـل مهمCRM هست.) هر مشـتری بـه دنبـال یـک نیـاز بـه دنبـال محصـولی میـرود و بعـد از یـافتن خواسـته هـایش در محصـول اقـدام بـه خریـد مـی کنـد. ولـی وفـاداری مشـتری کـه هـدف سیسـتم های CRM هست فقط از طریق جلب رضایت آنها به دست می آید.

فصل پنجم

ارتباط بین داده کاوی و مدریت ارتباط با مشتری

۱-۵ ارتباط بین داده کاوی و مدریت ارتباط با مشتری

اغلب تجارت ها به تصمیم گیری استراتژیک و یا اتخاذ خط مشی های جدید برای خدمت رسانی بهتر مشتریان نیاز دارند. به عنوان مثال فروشگاه ها آرایش مغازه خود را برای ایجاد میل بیشتر به خرید مجددا طراحی می کنند. این مثال به داده هایی در مورد رفتار مصرفی گذشته مشتریان برای تعیین الگویی به وسیله داده کاوی ، نیاز دارند.

برای روشن تر شدن مساله می توان مثال را این گونه بیان کرد که در یک فروشگاه زنجیره ای پس از داده کاوی مشخص می شود که درصدی از مشتریان خرید تلویزیون ، میز تلویزیون و گلدان کریستالی را هم در همان روز و بعد از خرید تلویزیون می خرند. مدیر فروشگاه می تواند بلافاصله دستوراتی صادر کند که بر اساس مدل های تلویزیون موجود میزهایی و براساس مدل میزها گلدان های کریستالی برایفروش سفارش داده شود و غرفه های جنبی غرفه تلویزیون را به میز و گلدان کریستالی اختصاص دهد . مطمئنا" حتی پس از مدت کوتاهی سود حاصل از این بخش از فروشگاه به طور قابل ملاحظه ای ترقی خواهد کرد .

در واقع ابزار داده کاوی ،داده را می گیرد و یک تصویر از واقعیت به شکل مدل می سازد که این مدل روابط موجود در داده ها را شرح می دهد. برای بهبودی بهره وری از یک فروشگاه داده کاوی از داده های انبار داده، مدل هایی را ارائه می دهد که بیانگر این هستند که چه محصولاتی یا خدماتی ،به چه مشتریانی،در چه زمانی واز طریق چه کانالی عرضه شود.

بیشتر شرکت ها بانک های داده ای عظیمی،شامل داده های بازاریابی،منابع انسانی و مالی را دارا هستند . بنابراین سرمایه گذاری در زمینه انبار داده ، یکی از اجزای حیاتی در استراتژی ارتباط با مشتری است.رابطه مشتری با زمان تغییر می کند و چنانچه تجارت و مشتری درباره یکدیگر بیشتر بدانند این رابطه تکامل ورشد 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 مشتری می گوید چه چیزی احتمالا" جالب توجه است و چه تصمیماتی باید گرفته شود. داده کاوی می تواند سود آوری مشتری های بالقوه را که می توانند به مشتریان بالفعل تبدیل شوند،پیش بینی کند و اینکه مشتری تا چه زمانی وفادار خواهد ماند وچگونه احتمالا" مارا ترک خواهد کرد.بعضی از مشتریان مرتبا" مراجعاتشان رابه شرکت ها برای کسب مزیت هایی که طی رقابت میان آنها به وجود می آید، تغییر می دهند.

در این صورت شرکت ها می توانند هدفشان را روی مشتریانی متمرکز کنند که سود آوری بیشتری دارند .

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

نتیجه گیری

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

منابع و مراجع

1- کتاب سال شبکه (کتاب مرجع صنعت شبکه در ایران) اذر ماه سال 1390 مقاله دوم peer1 hosting

2- کتاب مدریت کیفیت فراگیر (جلد اول) دکتر مصطفی جعفری موسسه خدمات فرهنگی رسا

3- مقاله داده کاوی ،مهندس مهرداد اشکاوندی راد

4- Fayyad U., Piatetsky-Shapiro G., and Smyth P., "From Data Mining to Knowledge Discovery in Databases," American Association for Artificial Intelligence, 1996.

5- Introduction to Data Mining and Knowledge Discovery By Two Crows Corporation

6- Top 10 algorithms in data mining XindongWu Knowl Inf Syst (2008)

7- سایت http://www.wikipedia.org

8- مقاله مدریت ارتباط با مشتری ، پرستو شاه محمدی

9- سایت http://irandataminer.ir در مورد وب کاوی

10- پایان نامه خانم شیما شاه حسینی ،دانشگاه علم وصنعت ،وب کاوی،1389

11- سایت www.modiriran.ir در رابطه با مدریت ارتباط با مشتری

12- مقاله اقای ارمان محمد زاده ،دانشگاه ازاد اسلامی واحد خمینی شهر، وب کاوی ،1391