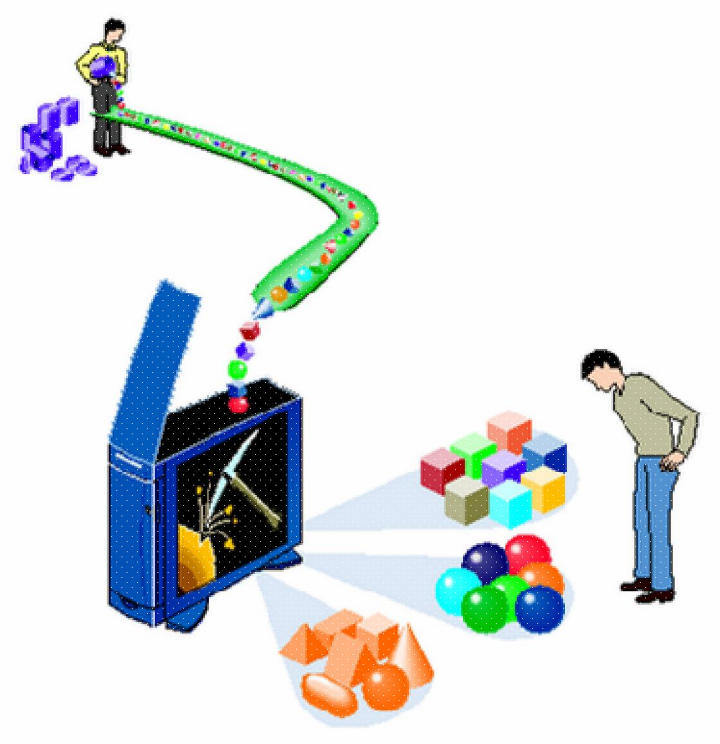
داده کاوی

DATA MINING

تهیه و تنظیم :

مهندس ایمان اشکاوند راد



1-1 - تاریخچه‌ی داده کاوی

با رشد فناوری اطلاعات و روش‌های تولید و جمع آوری داده ها، پایگاه داده های مربوط به داده های تبادلات تجاری، کشاورزی، اینترنت، جزئیات مکالمات تلفنی، داده های پزشکی و غیره سریعتر از هر روز جمع آوری و انبارش می شوند. لذا از اواخر دهه 80 میلادی بشر به فکر دست یابی به اطلاعات نهفته دراین پایگاه داده های حجیم افتاد زیرا سیستمهای سنتی قادر به این کار نبودند. به دلیل رقابت در عرصه‌های سیاسی، نظامی، اقتصادی ، علمی و اهمیت دست یابی به اطلاعات در کمترین زمان بدون دخالت انسان علم تجزیه و تحلیل داده ها یا داده کاوی پا به عرصه گذاشت.

داده کاوی[1] فرآیندی است که در آغاز دهۀ 90 مطرح شد و با نگرشی نو، به مسئلۀ استخراج اطلاعات از پایگاه داده ها می پردازد. از سال 1995 داده کاوی به صورت جدی وارد مباحث آمار شد و در سال 1996، اولین شمارۀ مجلۀ کشف دانش و معرفت از پایگاه داده ها[2] منتشر شد. محققانی نظیر براچمن و آناند (1996) کلیه مراحل واقع گرایانه و رو به جلو کشف دانش از پایگاه داده ها را تشخیص دادند.

در حال حاضر، داده کاوی مهمترین فناوری جهت بهره برداری موثر از داده های حجیم است و اهمیت آن رو به فزونی است. به طوریکه تخمین زده شده است که مقدار داده ها در جهان هر20 ماه به حدود دو برابر می رسد. در یک تحقیق که بر روی گروه های تجاری بسیار بزرگ در جمع آوری داده ها صورت گرفت مشخص گردید که 19 درصد از این گروه ها دارای پایگاه داده هایی با سطح بیشتر از 50 گیگا بایت می باشند و 59 درصد از آنها انتظار دارند که در آینده ای نزدیک در چنین سطحی قرار گیرند. [1]

درصنایعی مانند کارت های اعتباری و ارتباطات و فرشگاه های زنجیره ای و خریدهای الکترونیکی و اسکنرهای بارکد خوان هر روزه داده های زیادی تولید و ذخیره می شوند. افزایش سرعت کامپیوترها باعث به وجود آمدن الگوریتم هایی شده است که قدرت تجزیه و تحلیل بسیار بالایی دارند بدون اینکه محدودیتی در زمینه ظرفیت و سرعت کامپیوترها داشته باشند [2].

در سال 1989 و 1991 کارگاههای کشف دانش و معرفت از پایگاه داده ها توسط پیاتتسکی[3] و همکارانش برگزار شد. در فواصل سالهای 1991 تا 1994کارگاههای کشف دانش و معرفت از پایگاه داده ها توسط فییاد[4] و پیاتتسکی و دیگران برگزار شد. به طور رسمی اصطلاح داده کاوی برای اولین بار توسط فییاد در

[1] Data Mining
[2] Knowledge Discovery in Database(KDD)
[3] Piatetsky
[4] Fayyad

اولین کنفرانس بین المللی"کشف دانش و داده کاوی[1]" در سال 1995 مطرح شد. امروزه کنفرانسهای مختلفی در این زمینه در سراسر دنیا برگزار می شود.

افزایش داده های بسیار باعث پیدایش فرصتهای 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 و علوم کامپیوتر ظهور کرده است. مهندسی صنایع با حوزه های گوناگون و در برداشتن فرصتهای بی نظیر اکنون برای کاربرد داده کاوی و کشف دانش از پایگاه داده ها و برای توسعه مفاهیم و روشهای تازه در این زمینه آماده است. فرآیندهای صنعتی زیادی اکنون برای مطمئن شدن از کیفیت سفارشات محصول و کاهش هزینه های محصول به طور خودکار و کامپیوتری شده اند [3] .

1 - 2 - داده کاوی چیست؟

نگاهی به ترجمه تحت اللفظی داده کاوی، به ما در درک بهتر این واژه کمک می کند. Mine به معنای استخراج از منابع نهفته و با ارزش زمین اتلاق می شود. پیوند این کلمه با کلمه داده، جستجویی عمیق جهت پیدا کردن اطلاعات اضافی مفید که قبلاً نهفته بودند، از داده های قابل دسترس حجیم، را پیشنهاد می کند [4].

داده کاوی یک رشته نسبتاً جدید علمی می باشد که از انجام تحقیقات در رشته های آمار، یادگیری ماشین، علوم کامپیوترخصوصاً مدیریت پایگاه داده ها شکل گرفته است [5].

تعاریف متنوعی از داده کاوی در مراجع مختلف و توسط افراد مختلف ارائه شده است. از جمله:

1. داده کاوی عبارت است از فرآیند 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.[6]
2. اصطلاح داده کاوی به فرآیند نیمه خودکار تجزیه و تحلیل پایگاه داده های بزرگ به منظور یافتن الگوهای مفید اطلاق می شود [7].
3. داده کاوی یعنی جستجو در یک پایگاه داده ها برای یافتن الگوهایی میان داده ها [8].
4. داده کاوی یعنی تجزیه و تحلیل مجموعه داده های قابل مشاهده برای یافتن روابط مطمئن بین داده ها.
5. عبارت داده کاوی مترادف با یکی از عبارت های استخراج دانش، برداشت اطلاعات، وارسی داده ها و حتی لایروبی کردن داده هاست که در حقیقت کشف دانش در پایگاه داده ها (KDD) را توصیف می کند[7]

[1] Knowledge Discovery and Data Mining

اما تعریفی که در اکثر مراجع به اشتراک ذکر شده عبارت است از "استخراج اطلاعات و دانش و کشف الگوهای پنهان از پایگاه داده های بسیار بزرگ و پیچیده ". داده کاوی یک متدولوژی بسیار قوی و با پتانسیل بالا می باشد که به سازمان ها کمک می کند که بر روی مهمترین اطلاعات از مخزن داده های خود تمرکز نمایند. [1]

داده کاوی فرآیندی است که از ابزارهای تحلیلی گوناگونی برای کشف الگوها و روابط بین داده ها استفاده می کند که ممکن است برای اعتبار بخشیدن به پیش بینی استفاده شود [10].

داده کاوی کمک می کند تا سازمان ها با کاوش بر روی داده های یک سیستم، الگوها و رفتارهای آینده را کشف و پیش بینی کرده و بهتر تصمیم بگیرند. داده کاوی با استفاده از تحلیل وقایع گذشته یک تحلیل اتوماتیک و پیش بینانه ارائه می نماید و به سوالاتی جواب می دهد که پاسخ آنها در گذشته ممکن نبوده و یا به زمان زیادی نیاز داشته است.

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

1 - 3 - مراحل داده کاوی

داده کاوی در این چرخه خود نیز شامل مراحل مختلفی می باشدکه عبارتند از:

1. تعیین اطلاعات گذشته.

2. تمیز کردن داده ها و پردازش اولیه. دراین مرحله خطاهای داده ها تصحیح می شوند و داده های اشتباه جایگزین می شوند. این مرحله ممکن است تا 60 درصد از زمان داده کاوی را در برگیرد [1]

3. یکپارچه سازی داده ها. معمولاً داده ها از منابع متفاوتی جمع آوری می شوند باید به صورتی درآیند که یک مخزن از داده های[1] مناسب ایجاد شود تا بتوان عملیات داده کاوی را بهتر انجام داد.

4. انتخاب مجموعه داده های هدف.

5. یافتن ویژگیهای مورد استفاده و تعیین ویژگی های جدید.

6. نمایش داده ها به صورتی که بتوان برای داده کاوی استفاده نمود.

[1] Data Warehouse

7. انتخاب عملیات داده کاوی (طبقه بندی، خوشه بندی، پیش بینی وغیره).

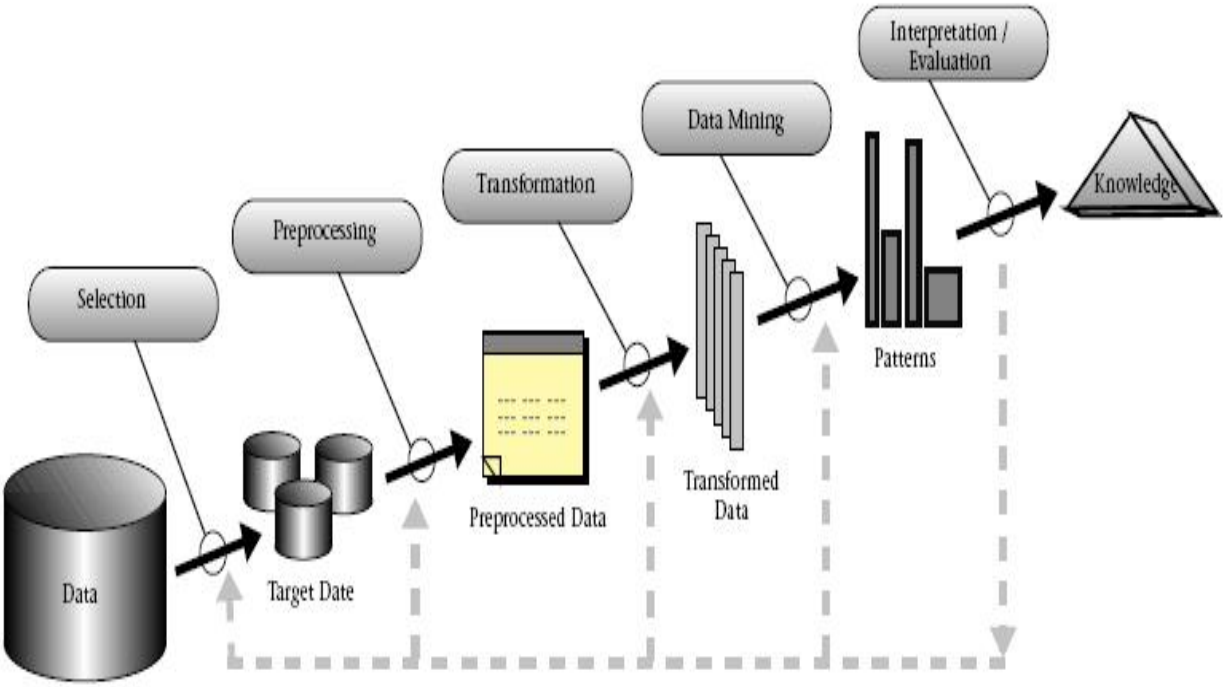
8. انتخاب روش داده کاوی (شبکه های عصبی، درخت تصمیم و نظایر آن).

9. داده کاوی و جستجو برای یافتن الگوی مناسب.

10 . ارزیابی و تحلیل الگوی به دست آمده و حذف الگو های نامناسب.

11 . تفسیر نتایج داده ها و استنتاج از اطلاعات با ارزش.

باید توجه داشت که جمع آوری و محافظت از داده ها نکته بسیار مهمی می باشد. اصولاً چون قالب و نوع داده ها در طول زمان تغییر می کند ممکن است بسیاری از داده های موجود در قالبهای متفاوت باشند و همچنین بسیاری از داده های قدیمی از بین رفته و دور ریخته شوند. در حالی که ممکن است اهمیت این داده ها از داده های جدید به هیچ وجه کمتر نباشد. همچنین به علت این که داده ها می توانند از منابع مختلف داخلی و خارجی مانند کارکنان شرکت، مدیران، مشتریان، کارفرمایان، پیمانکاران باشند باز هم ممکن است قالب داده ها با هم یکسان نباشد. به همین دلیل انتخاب داده های درست و یکپارچه سازی قالب آن ها به منظور استفاده در داده کاوی از اهمیت بسیار بالایی برخوردار می باشد. در شکل 1-1 می‌توان مراحل داده کاوی را به اختصار نشان داد [11].



شکل 1-1 مراحل داده‌کاوی

1 - 4 - برخی از کاربردهای داده کاوی در محیطهای واقعی

1. **خرده فروشی** : از کاربردهای کلاسیک داده کاوی است که می توان به موارد زیر اشاره کرد :
   - تعیین الگوهای خرید مشتریان
   - تجزیه و تحلیل سبد خرید بازار
   - پیشگویی میزان خرید مشتریان از طریق پست(فروش الکترونیکی)
2. **بانکداری** :
   - پیش بینی الگوهای کلاهبرداری از طریق کارتهای اعتباری
   - تشخیص مشتریان ثابت
   - تعیین میزان استفاده از کارتهای اعتباری بر اساس گروههای اجتماعی
3. **بیمه** :
   - تجزیه و تحلیل دعاوی
   - پیشگویی میزان خرید بیمه نامه های جدید توسط مشتریان
4. **پزشکی** :
   - تعیین نوع رفتار با بیماران و پیشگویی میزان موفقیت اعمال جراحی
   - تعیین میزان موفقیت روشهای درمانی در برخورد با بیماریهای سخت

1 - 5 - عملیات داده کاوی

مجموعه عملیاتی را که روش داده کاوی قادر به انجام آن است در ذیل به صورت کامل تشریح شده اند.

1 - 5 - 1 - طبقه بندی و پیشگویی[1]

طبقه بندی یکی از عملیات رایج و مورد استفاده در داده کاوی است. طبقه بندی عملیاتی است که سازمانها را قادر می سازد که در حل مسائل خاص در مجموعه های بزرگ و پیچیده به کشف الگو ها دست یابند. طبقه بندی فرآیندی می باشد که مجموعه داده ها را به قسمت های مشخص تقسیم می کند. برای مثال مشتریان یک شرکت بیمه را بر اساس خصوصیاتشان به دو گروه با ریسک بالا و ریسک پایین تقسیم می‌کند. با این کار در واقع مشتریان این شرکت طبقه بندی شده اند.

[1] Classification and Prediction

ساده ترین روشی که برای طبقه بندی به نظر می رسد گذاشتن حدی برای دسته ها می باشد، مثلاً افـراد بـا درآمد بالای مقداری مشخص را به یک دسته و افراد بـا درآمـد پـایین تـر از آن را بـه یـک دسـته ی دیگـر تخصیص دهیم.

میشل (1997)، مولر و چرکاسکی(1998)، تعدادی از روشـهایی کـه مـی تواننـد جهـت داده کـاوی مسـائل طبقه‌بندی به کار برده شوند، شامل : درخت تصمیم و شبکه های عصبی و نظیر این ها را ارایـه کردنـد. ایـن روشها در دامنه گسترده‌ای از زمینه های مهندسی به کار برده می شوند. برای نمونه، شبکه هـای عصـبی در کنترل بازخورد ها برای کشف الگوها و آشکارسازی خروجی مناسب کنترل شده به کار برده می شوند.

طبقه بندی داده ها یک فرآیند دو مرحله ای می باشد. در گام اول، یک مدل بر اسـاس مجموعـه داده هـای آموزشی موجود در پایگاه داده ها ساخته می گردد. مجموعه داده های آموزشی از رکوردها، نمونه هـا، مثالهـا و یا اشیائی که شامل مجموعه‌ای از صفات یا جنبه ها می باشد، تشکیل شده انـد. هـر نمونـه یـک برچسـب کلاس معلوم دارد، که در یکی از صفات به نام برچسب کلاس مشخص شده است. بـه هریـک از نمونـه هـای مجموعه داده های آموزشی، یک نمونهٔ آموزشی گویند، کـه بـه طـور تصـادفی از مجموعـه داده هـا انتخـاب می‌شود. زمانی که برچسب کلاس آموزشی مشخص باشد، این مرحله از یـادگیری را یـادگیری نظـارت شـده (یادگیری با ناظر) می نامند. نوع دیگری از یادگیری بدون نظارت (یادگیری بدون ناظر) می باشد، کـه در آن برچسب کلاس هر نمونه آموزشی نامعلوم است (مانند: خوشه بندی). به طور معمول، مـدلهای سـاخته شـده به فرمهایی از قواعد طبقه بندی و درخت تصمیم نشان داده می شوند.

به عنوان مثال یک پایگاه داده ها شـامل اطلاعـات مشـتریان کارتهـای اعتبـاری را در نظـر بگیریـد، قواعـد طبقه‌بندی می تواند جهت طبقه بندی مشتریان به نرخ اعتباری عالی و خـوب سـاخته شـوند. از ایـن قواعـد می‌توان جهت طبقه بندی نمونه داده های جدید استفاده کرد.

در گام دوم مدل برای طبقه بندی مناسب مشتریان جدید استفاده می شود. قواعـد یـادگیری کـه از تحلیـل داده های مشتریان موجود حاصل شده است، می تواند برای پیشگویی کلاس اعتبار مشتریان جدید یا آینـده مورد استفاده قرار گیرد.

از نقطه نظر کلی، طبقه بندی و رگرسیون دو نوع اصلی از مسائل پیشگویی هستند، که طبقـه بنـدی جهـت پیشگویی مقادیر گسسته و اسمی مورد استفاده قرار می گیرد، در حالی که رگرسیون جهت پیشگویی مقادیر پیوسته مورد استفاده قرار می گیرد. در اینجا مـا پیشـگویی را بـرای پیشـگویی برچسـب کـلاس بـه عنـوان طبقه‌بندی و برای پیشگویی مقادیر پیوسته، به عنوان پیشگویی معرفی می کنیم.

طبقه بندی و پیشگویی کاربردهای زیادی در بازرگانی، بانکداری، پزشکی، ارتباطات، کشاورزی و غیره دارد.

طبقه بندی را می توان به عنوان یک فرایند دو مرحله ای در نظر گرفت. اول، یک مدل طبقه بندی با توجه به مجموعه داده های آموزشی ساخته می شود. چنین مدلی می تواند به فراهم کردن یک درک بهتر از داده‌های گمشده کمک کند. به طور معمول، این مدلها به فرمهایی از درخت تصمیم، یا فرمولهای ریاضی نمایش داده می شود. سپس مدل می تواند قوانین اگر-آنگاه را جهت پیشگویی برچسب های کلاس داده‌های جدید که دارای برچسب کلاس نامعلوم هستند، مورد استفاده قرار دهد.

1 - 5 - 1 - 1 - روشهای طبقه بندی

روشهای طبقه بندی در داده کاوی عبارتند از:

1- رگرسیون خطی چند گانه

2- رگرسیون لجستیک

3- تحلیل ممیزی

4- بیز ساده

5- شبکه های عصبی

6- درختهای تصمیم

7- K - نزدیکترین همسایگی

1 - 5 - 2 - خوشه بندی

خوشه بندی یکی از مهمترین ابزار کشف داده ها است که در کشف های تصادفی به کار گرفته می شود. در حال حاضر، اخذ دانش یک گلوگاه عمده در فرآیند مهندسی دانش محسوب می شود. الگوریتم های یادگیری ماشین و داده کاوی با هدف استخراج دانش از داده ها، به عنوان روشی برای حل این مشکل مطرح می باشند. یک رهیافت متداول در این زمینه روش خوشه بندی است که برای تصمیم گیری یا طبقه بندی یا کلاس بندی می تواند تصمیمات نمادینی را به نمونه های جدید با استفاده از نمونه های موجود منتسب کنند. روش های خوشه بندی به واسطه قابلیت درکی که در خود نهفته دارند، از اقبال خوبی برخوردار شده‌اند. وجود قابلیت درک از جهات گوناگونی حائز اهمیت می باشد: فهم قلمرو، درک قابلیت های

کلاس‌بندی، توجیه تصمیم و بالاخره وجود قوانین نمادینی که می توانند از روی خوشه های استخراج شده و سپس در یک سیستم تصمیم گیری مبنی بر قوانین به کار گرفته شوند.

خوشه بندی در واقع یک عملیات غیر نظارتی می باشد. این عملیات هنگامی استفاده می شود که ما به دنبال یافتن گروه هایی از داده های مشابه می باشیم بدون اینکه از قبل پیش بینی در مورد شباهت های موجود داشته باشیم. خوشه بندی معمولاً هنگامی استفاده می شود که به دنبال یافتن گروه هایی از مشتریان هستیم که قبلاً شناخته نشده اند. برای مثال می توان شباهت های مشتریان در استفاده از تلفن همراه را به منظور گروه بندی مشتریان و تشخیص خدمت جدیدی جستجو نمود.

خوشه بندی عملی است که در طی آن گروههایی از داده ها و یا اقلام وجود دارند به طوری که هر مورد به یک خوشه نسبت داده می شوند و اعضای داخل خوشه نیز باید دارای شباهت ذاتی با هم باشند و معیار اندازه گیری شباهت باید کاملاً مشخص باشد و برای هر جفت از موارد قابل محاسبه باشد. بنابراین در هر خوشه یک خود شباهتی[1] بین اقلام آن خوشه وجود دارد.

پایگاه های داده بسیار بزرگ ممکن است شامل متغیر های بسیار زیاد، ابعاد بسیار بزرگ و ساختار بسیار پیچیده باشند به طوریکه حتی بهترین روش های داده کاوی مستقیم هم نمی توانند الگو های معنی داری در آن ها را استخراج نمایند. در خیلی از موارد مشکل این نیست که الگویی برای کشف شدن وجود ندارد بلکه در واقع تعداد زیادی الگو وجود دارد ولی روشهای داده کاوی برای جواب دادن به سوالی که مطرح شده است، الگویی کشف نمی کنند.

در بازاریابی ممکن است افراد، جامعه را به وسیله متغیر هایی که از قبل به عنوان معیارهای مناسبی می‌شناختیم طبقه بندی نماییم. در حالی که ممکن است به دلیل پیچیدگی پایگاه داده ها نظری در مورد متغیر های طبقه بندی کننده و یا چگونگی تعیین و یا خوشه ها نداشته باشیم. دراین گونه موارد است که به سراغ روشهای خوشه بندی می رویم [5].

خوشه بندی یک روش داده کاوی غیر مستقیم است. برای اکثر روش های داده کاوی مثل درخت تصمیم گیری و شبکه های عصبی، با یک مجموعه آموزشی شروع کرده و به کمک این مجموعه سعی می شود یک مدل برای بخش بندی داده ها ، ایجاد گردد. سپس از آن مدل برای پیش بینی داده های جدید استفاده شود.

در روش خوشه بندی هیچ دسته ای از قبل وجود ندارد و در واقع متغیر ها به صورت مستقل و وابسته تقسیم نمی شوند. بلکه ما در اینجا به دنبال گروهها یی از داده ها هستیم که به هم شباهت دارند و با کشف

[1] Self Similarity

این شباهت ها می توان رفتارها را بهتر شناسایی کرد و بر مبنای آنها طوری عمل کرد که نتیجه بهتری حاصل شود.

1 - 5 - 3 - تحلیل روابط و وابستگیها[1]

پیشرفت تکنولوژی فروشگاه های خرده فروشی را قادرساخته است حجم زیادی از داد ههای مربوط به خر ید هر یک از مشتریان که از آن به عنوان سبد بازار یاد می شود را جمع آوری و ذخیره نمایند. فراهم بودن جزئیات اطلاعات ثبت شده مشتریان منجر به بهبود روش هایی شده است که به طور اتوماتیک روابط بین آیتم هایی که در پایگاه داده ها انبارش شده اند را جستجو می کنند.

همزمان با پیدایش علم داده کاوی در اوایل دهه 90 الگوریتم های استخراج قوانین وابستگی از پایگاه داده ها نیز پا به عرصه گذاشت. نویسندگان زیادی در زمینه استخراج قوانین وابستگی در پایگاه داد ه ها بحث کرده‌اند. در [4] به مقایسه‌ی الگوریتمهای مهم استخراج قوانین وابستگی، مزیت ها و معایب الگوریتم‌ها پرداخته شده است.

اساساً ارتباط میان مجموعه اشیاء وابستگی های جالب توجهی هستند که منجر به امکان آشکار سازی الگوهای مفید و قوانین وابستگی برای پشتیبانی تصمیم، پیش بینی های مالی، سیاست های بازاریابی، وقایع پزشکی و خیلی کاربرد های دیگر می شود. در حقیقت توجهات زیادی را در تحقیقات اخیر به خود جلب کرده است [4].

تحلیل وابستگی ها یک حالت غیر نظارتی داده کاوی می باشد که به جستجو برای یافتن ارتباط در مجموعه داده ها می پردازد .یکی از کاربردی ترین حالات تحلیل وابستگی ها " تجزیه تحلیل سبد بازار" می باشد که در آن هدف یافتن کالا هایی است که معمولاً به طور همزمان خریداری می شوند. این کار کمک می کند که خرده فروشان بهتر بتوانند کالاهای خود را سازماندهی کرده و چیدمان بهتری از محصولات خود داشته باشند [10].

داده های موجود در سبد بازار نشان دهنده خرید مشتری دریک زمان خاص هستند. هر مشتری خرید مجزایی را در کمیته های مختلف و زمانهای متفاوت انجام می دهد. با تجزیه و تحلیل سبد بازار بینشی برای خرده فروشان از اینکه چه محصولاتی با هم خریداری می شوند فراهم می‌گردد و بنابراین می توانند رفتارخرید مشتریان را پیش بینی کنند این کار به آنها کمک می کند که بهتر بتوانند کا لاهای خود را

[1] Association Analysis

سازماندهی کرده و چیدمان بهتری از محصولات خود داشته باشند و بنابراین سودآوری خود را افزایش دهند [10].

1 - 5 - 4 - پیش بینی[1]

در طبقه بندی گروه هایی مشخص می شوند که اقلام به آن ها تعلق دارند. پیشگویی هایی که براساس مدلهای طبقه بندی ارایه می شوند دارای یک خروجی گسسته می باشد که مشخص می کند که مثلاً یک مشتری جزء گروه با پاسخ مثبت است یا منفی و یک مریض جزء گروه با ریسک بالا است یا پایین. ولی پیش بینی بر خلاف پیش گویی[2] یک مقدار پیوسته را پیش بینی می کند مثلاً تقاضای آینده با قیمت نفت در سال آینده. پیش بینی معمولاً به وسیله رگرسیون (عملیاتی که با تعیین ارتباط بین متغیر ها به پیش بینی می‌پردازد) صورت می گیرد. بسته های نرم افزاری مانند SAS و SPSS معمولاً توانایی حل مساله های پیچیده را فراهم می‌نمایند. ولی استفاده از چنین عملیات آماری نیاز به دانش بالای آمار در خصوص شرایط و چگونگی استفاده از این ابزارها را دارد. ابزارهای داده کاوی نظیر شبکه های عصبی نیز به وفور برای پیش‌بینی استفاده می شود.

از مسایل ساده پیش بینی عبارتند از : پیش بینی مقادیر پیوسته بر اساس یکسری داده های موجود. برای مثال پیش بینی درآمد یک فرد بر اساس مشخصات فرد. ابزارهایی نظیر درخت تصمیم گیری و شبکه های عصبی چنین کاری را انجام می دهند.

از مسایل پیچیده پیش بینی می توان به پیش بینی یک یا چند مقدار براساس الگوهای تکراری و متوالی مانند سطح سهام بازار در 30 روز آینده بر اساس داده های 6 ماه گذشته اشاره کرد. ابزارهای داده کاوی به سختی چنین پیش بینی هایی را انجام می دهند. در این گونه مواقع داده های موجود باید به صورتی مناسب و در جهت مناسب استفاده شوند و فرمت داده های خروجی به درستی مشخص باشد. همچنین در این گونه پیش بینی ها نیاز به یک تحلیلگر به منظور پردازش داده های ورودی و تحلیل داده های خروجی بیشتر احساس می‌شود.

1 - 6 - زیربنای داده کاوی

تکنیکهای داده کاوی نتیجه ی تحقیقات گسترده و بلند مدتی است که در طول سالها برای افزایش بازدهی تجاری موسسات بکار برده می‌شدند. تحقیقات در این زمینه از زمانی آغاز شد که برای نخستین بار اطلاعات تجاری هر سازمان، بر روی سیستمهای ذخیره سازی آن زمان که از نوع مغناطیسی بودند، ذخیره

[1] Forecasting
[2] Prediction

شدند. این رشته تحقیقات با توسعه و پیشرفت سیستمهای اطلاعات که قابلیت ذخیره ی حجم بیشتری از داده ها را فراهم می‌کردند و همچنین از سرعت بسیار بالاتری در ذخیره سازی و بازیابی اطلاعات برخوردار بودند، اهمیت بیشتری یافت. روشهای دسترسی تصادفی یا رندم به اطلاعات و پیدایش روشهای حرکت[1] در میان داده ها، خصوصاً بصورت بلادرنگ، فناوری داده کاوی را متحول ساخت.

روشهای داده کاوی بر پایه های زیر استوار هستند:

- گردآوری حجم عظیمی داده.
- کامپیوترهای چند پردازنده ی قدرتمند.
- الگوریتمهای داده کاوی.

در سالهای 1960 صنعت گردآوری اطلاعات و امکان ذخیره ی داده ها در تجهیزاتی نظیر نوار و دیسک توسط شرکتهایی که IBM و CDC از پیشگامان آنها بودند، شکل تجاری به خود گرفت. با رواج چنین مکانیسمهایی تبادل استاتیک اطلاعات امکانپذیر شده، پرسشهای تجاری از قبیل آنکه "سود خالص شرکت در پنج سال آخر فعالیت چقدر بوده است؟ " پاسخ داده می شود. 20 سال بعد از فناوری فوق، با پیشرفتهای نرم افزاری و استفاده از بانکهای اطلاعاتی رابطه ای[2] و زبان جستجوی ساخت یافته[3] توسط شرکتهای موفقی همچون MICROSOFT, IBM, INFORMIX, SYBASE, ORACLE، ... اطلاعات در همان لحظه ی ثبت شدن قابل تبادل بودند. بعبارت دیگر تبادل اطلاعات بصورت دینامیک امکانپذیر شده بود. نمونه ای از سوالات تجاری که این سیستم پاسخگوی آن است چنین بود: "مقدار فروش شعب [کشور یا شهر مورد نظر] در ماه مارس گذشته چه میزان بوده است؟". در سالهای دهه ی نود نوبت به تکنولوژی هایی همچون انبار داده ها[4] و امکانات تصمیم گیری نرم افزاری رسید. [9]

1 - 7 - تکنولوژی های مرتبط با داده کاوی

1- پردازش تحلیل روی خط[5] - OLAP
2- بانکهای اطلاعاتی چند بعدی[6]
3- انبار داده ها

پیشگامان ابزارهای نرم افزاری چنین تکنولوژیهایی شرکتهایی نظیر Pilot, Comshare, Arbor Cognos, Microstrategy بودند. البته بلافاصله در همان زمان شرکتهایی نظیر ORACLE, IBM MICROSOFT، که امروزه نام آنها را در همه جا مشاهده می کنیم نیز کنترل جریان را بدست گرفته و نرم افزارهای آنها

[1] navigation
[2] RBDMS
[3] SQL
[4] DataWareHousing
[5] OnLine Analitical Process
[6] MultiDimensional Databases

بازار را تسخیر کرد. هسته ی فناوری داده کاوی شامل علوم آمار، هوش مصنوعی، آموزش ماشین و علوم نوین دیگری است که در طول سالهای گذشته پیشرفت قابل توجهی داشته است.

1 - 8 - چه نوع اطلاعاتی مناسب داده کاوی است؟

ما مقادیر انبوهی از اطلاعات از داده‌های عددی ساده و سندهای متنی تا اطلاعات پیچیده ای همچون داده‌های چند بعدی، فایهای چندرسانه‌ای و اسناد ابر متن را جمع‌آوری می‌کنیم. در زیر لیستی از گونه‌های مختلف جمع‌آوری شده در قالب فرمهای دیجیتالی در پایگاههای داده و فایل‌های ساده‌ی متنی آمده است.

1 - 8 - 1 - مبادلات و تراکنشهای تجاری

معمولاً تمامی مبادلات و تراکنشهای صنعتی و تجاری بصورت دائمی ذخیره و نگهداری می‌شوند. چنین مبادلاتی معمولاً وابسته به زمان بوده و شامل تعاملات بین‌التجار مثل خریدها، تعویضها، بانکداری، سهام و ... بوده و یا شامل کنش‌های متقابل تجاری مانند مدیریت کالاها و وسایل خانه می‌باشد. برای نمونه فروشگاههای زنجیره‌ای بزرگ به لطف کاربرد فراگیر بارکدها، روزانه میلیونها تراکنش را در قالب چندین ترابایت داده، ذخیره و نگهداری می‌کنند. مشکل اصلی، فضای ذخیره‌سازی این حجم داده نیست، چرا که قیمت رسانه‌های ذخیره‌سازی روز به روز در حال کاهش است. در واقع بکارگیری موثر این قبیل داده‌های جمع‌آوری‌شده، آن هم در یک بازه‌ی زمانی مناسب، برای تصمیم گیری در بازار رقابتی امروز، مهمترین مشکل برای حل مشکلات تصمیم گیری و نجات پیدا کردن از این دنیای رقابتی می‌باشد.

1 - 8 - 2 - داده‌های علمی

چه در لابراتوار شمارش ذرات شتاب دهنده‌ی هسته‌ای در سوئیس، چه در مطالعه‌ی اطلاعات رادیویی حاصله از قلاده‌های خرسهای گریزلی در کانادا، چه در جمع‌آوری اطلاعات در مورد فعل و انفعالات اقیانوسی از کوههای شناور در قطب جنوب و چه در روانشناسی روی انسانها در یک دانشگاه امریکایی، جامعه‌ی ما در حال جمع‌آوری مقادیر بسیار زیادی اطلاعات علمی است که نیاز به پردازش و تجزیه و تحلیل دارند، متاسفانه میتوان اطلاعات بسیار مفیدی را از داده‌های کهنه شده‌ای که هنوز کاملاً جمع‌آوری نشده‌اند، استخراج و نگهداری کرد، بسیار سریعتر از آنکه بخواهیم داده‌های قدیمی و منقضی شده ای را جمع‌آوری و سپس مورد تجزیه و تحلیل قرار دهیم.

1 - 8 - 3 - داده‌های بهداشتی و شخصی

از سرشماریهای دولتی گرفته تا فایل‌های افراد و مشتریان، مجموعه‌هایی از اطلاعات بطور پیوسته در مورد اشخاص و گروهها در حال جمع‌آوری است. دولتها، کمپانی‌ها و سازمانهایی مثل بیمارستانها، مقادیر بسیار مهمی از اطلاعات شخصی را برای کمک در مدیریت منابع انسانی جمع آوری و انبار می‌کنند، همچنین برای درک بهتر از بازار و کمک و راهنمایی ساده‌تر مشتری، بدون توجه به سیاستهای صادره و

گزارش شده، اینگونه داده‌ها اغلب فاش می‌شوند و در موارد بسیاری گسترش می‌یابند. این گونه داده‌ها زمانی که بهمراه اطلاعات مهم دیگری گسترش یابند، ممکن است منجر به تغییر در سلیقه و رفتار مشتری شوند که تحلیل آنها اطلاعات بسیار مناسبی را در اختیار قرار می دهد.

1 - 8 - 4 - نظارت تصویری و ویدئویی

با افت قیمت شگفت انگیز دوربینهای تصویربرداری، استفاده از آنها بسیار فراگیر شده است. نوارهای ویدئویی دوربینهای امنیتی معمولاً بازیابی شده و اطلاعات قدیمی آنها از بین می رود،اگر چه امروزه گرایش بیشتر به نگهداری نوارها و حتی دیجیتالی کردن آنهاست.

1 - 8 - 5 - دریافتها و مشاهدات ماهواره ای

امروز تعداد غیر قابل شمارشی ماهواره گرداگرد جهان قرار دارد، که برخی از آنها ایستگاههای ناحیه ای بالای سطح زمین هستند و برخی دیگر در مداری گرد زمین می چرخند. به هر صورت تمامی آنها در حال ارسال جریان بدون وقفه ای از اطلاعات به سطح زمین هستند. NASA که کنترل تعداد زیادی از این ماهواره ها را در اختیار دارد، در هر ثانیه مقادیر زیادی از اطلاعات را، بیش از آنچه که تمامی مهندسین و محققین NASA می توانند جمع آوری کنند، دریافت می دارد. تعداد زیادی از تصاویر ماهواره ای و اطلاعات بمحض دریافت، پخش عمومی شده و در اختیار همگان قرار می گیرد، به امید آنکه سایر محققان بتوانند آنها را تجزیه و تحلیل نمایند.

1 - 8 - 6 - بازیهای المپیک

جامعه ی ما مقادیر زیادی اطلاعات و آمارها در مورد بازیهای المپیک، بازیکنان و ورزشکاران جمع آوری می نماید، از امتیازات هاکی و پاسهای بازیهای بسکتبال و تعداد دورهای طی شده در یک مسابقه ی رالی اتومبیلرانی گرفته، تا رکورد های شناگران، ضربات بکسرها و موقعیت مهره ها در بازیهای شطرنج، همه ی اینگونه اطلاعات جمع آوری می شوند. مفسرین و خبرنگاران از این اطلاعات برای گزارش وقایع استفاده می کنند، اما مربیان این اطلاعات را در جهت افزایش توان و نیرو و درک بهتر حریفان و رقیبان بکار می گیرند.

1 - 8 - 7 - رسانه ی دیجیتال

گسترش اسکنرهای ارزان قیمت، دوربینهای ویدئویی رومیزی و دوربینهای دیجیتال، یکی از دلایل گسترش تولیدات این رسانه هاست. بعلاوه بسیاری از ایستگاههای رادیویی، کانالهای تلویزیونی و استودیوهای فیلمسازی مشغول دیجیتال کردن مجموعه های ویدئویی و صوتی خود برای ارتقا در سطح مدیریت دارائی های چندرسانه ایشان می باشند. شرکتهایی همچون NHL و NBA فرآیند تبدیل مجموعه های عظیم بازی هایشان به فرمتهای دیجیتالی را آغاز کرده اند.

1 - 8 - 8 - داده های مهندسی نرم افزار و طراحی به کمک کامپیوتر[1]

سیستم های نرم افزاری متنوعی جهت طراحی به کمک کامپیوتر و جهت طراحی ساختمانها یا برای مهندسان جهت درک بهتر اجزای سیستم و مدارات وجود دارد. اینگونه سیستمها مقادیر نامتناهی داده تولید می کنند. علاوه بر این مهندسی نرم افزار، منبع مشابه قابل توجهی از داده در قالب کد، توابع کتابخانه ای، اشیا و ... می باشد که به ابزارهای قوی برای مدیریت و نگهداری نیازمند می باشد.

1 - 8 - 9 - دنیاهای مجازی

برنامه های کاربردی زیادی وجود دارند که از فضاهای مجازی سه بعدی استفاده می کنند. این فضاها و اشیایی که دارند با زبانهای ویژه ای همچون VRML تشریح می شوند. در حالت مطلوب این فضاهای مجازی به شیوه ای تشریح می شوند که میتوانند اشیا و فضاها را به اشتراک گذارند. در حال حاضر مقادیر قابل توجهی از اشیای مجازی و فضاهای ساخته شده موجود می باشد. مدیریت این منابع جمع آوری شده مانند جستجوی بر اساس محتوا و بازیابی از این مجموعه ها در حال گسترش و رشد است.

1 - 8 - 10 - گزارشات متنی و نامه های الکترونیکی

اکثر ارتباطات داخلی و بینابین شرکتها یا سازمانهای تحقیقاتی یا حتی اشخاص بر مبنای گزارشات و یادداشتهای در قالب متن بوده و اغلب این تبادلات با پست الکترونیکی انجام می شود. این پیغامها مرتباً در فرمها و قالبهای دیجیتالی برای کاربرد های آینده و همچنین ایجاد منابع و کتابخانه های عظیم دیجیتالی، نگهداری می شوند.

1 - 8 - 11 - منابع و اطلاعات موجود در شبکه ی جهانی وب

از زمان آغاز به کار شبکه ی جهانی وب در سال 1993، اسنادی از گونه ها و غالبهای مختلف، محتویات و جزئیات جمع آوری شده و مرتبط شده از داخل، با ابر پیوند ها آن را تبدیل به بزرگترین منبع داده ای کرده که تاکنون ساخته شده. بر خلاف طبیعت غیر ساخت یافته و دینامیک آن، خصوصیات نامتجانس، افزونگی و تناقضات زیاد موجود در آن، همچنین بدلیل تنوع وسیع آن و موضوعات پوشش داده شده، همچنین سهم بیکران آن از منابع و انتشارات، مهمترین مرجع داده ای است که تاکنون مورد استفاده عموم قرار گرفته. نظریات مختلف بر این اعتقادند که شبکه ی جهانی وب تالیفی از علوم بشر خواهد بود[12].

1 - 9 - انبار داده[2] چیست؟

انبار داده به مجموعه ای از داده ها گفته می شود که از منابع مختلف اطلاعاتی سازمان جمع‌آوری، دسته بندی و ذخیره می شود. در واقع یک انبار داده مخزن اصلی کلیه داده های حال و گذشته یک سازمان

[1] CAD
[2] Data Warehousing

می باشد که برای همیشه جهت انجام عملیات گزارش گیری و آنالیز در دسترس مدیران می باشد. انبارهای داده حاوی داده هایی هستند که به مرور زمان از سیستم های عملیاتی آنلاین سازمان[1] (OLTP) استخراج می شوند، بنابراین سوابق کلیه اطلاعات و یا بخش عظیمی از آنها را می توان در انبار داده ها مشاهده نمود.

از آنجائیکه انجام عملیات آماری و گزارشات پیچیده دارای بارکاری بسیار سنگینی برای سرورهای پایگاه داده می باشند، وجود انبار داده سبب می گردد که اینگونه عملیات تاثیری بر فعالیت برنامه های کاربردی سازمان (OLTP) نداشته باشد.

همانگونه که پایگاه داده سیستمهای عملیاتی سازمان (برنامه های کاربردی) به گونه ای طراحی می شوند که انجام تغییر و حذف و اضافه داده به سرعت صورت پذیرد، در مقابل انبار داده ها دارای معماری ویژه ای می‌باشند که موجب تسریع انجام عملیات آماری و گزارش گیری می شود[2].(OLAP)

[1] Online Transactional Processing Systems
[2] Online Analytical Processing System

منابع و مراجع

[1] An Introduction to Data Mining: http://www.thearling.com/, retrieved on Mar 2, 2007.

[2] Data Mining: Efficient Data Exploration and Modeling: http://research.microsoft.com/dmx/DataMining/ , retrieved on Mar 2, 2007.

[3] Christine Gertisio and Alan Dussauchoy, "Knowledge Discovery from Industrial Data base", Journal of Intelligent Manufacturing, 15, 29-37, 2004.

[4] Cornelia Gyorödi, Robert Gyorödi, Stefan Holban – "A Comparative Study of Association Rules Mining Algorithms", SACI 2004, 1st Romanian-Hungarian Joint Symposium on Applied Computational Intelligence, Timisoara, Romania, May 25-26, pag. 213-222, 2004.

[5] Berson A., Smith S., and Thearling K., “Building Data Mining Applications for CRM,” Tata McGraw-Hill, New York, 2004.

[6] Introduction to Data Mining and Knowledge Discovery By Two Crows Corporation

[7] Jeffery W. Seifert , Analyst in information science and Technology Policy, ‘ Data Mining : An Overview ‘ December 2004.

[8] David J. HAND , Data Mining: Statistics and More? , December 2002.

[9] Jeffrey W. Seifert ,Data Mining:An Overview, CRS Report for Congress, December 16, 2004

[10] Berry, M., and Linoff, G., “Data Mining Techniques: For Marketing, Sales, and Customer Support” New York: John Wiley and Sons, 1997.

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

[12] http://www.exinfm.com/pdffiles/intro_dm.pdf

http://ashkavand.blogfa.com

Email: i81a@yahoo.com