معرفی داده کاوی یاData mining

معرفی داده کاوی یاdata mining داده کاوی یا کشف دانش در پایگاه داده ها (KDD) علم نسبتا تازه ای است که با توجه پیشرفت کشور در زمینه IT و نگاه های ویژه به دولت الکترونیک و نفوذ استفاده از سیستم های رایانه ای در صنعت و ایجاد بانک های اطلاعاتی بزرگ توسط ادارات دولتی، بانک ها و بخش خصوصی نیاز به استفاده از آن به طور عمیقی احساس می شود .

معرفی داده کاوی

داده کاوی یا کشف دانش در پایگاه داده ها (KDD) علم نسبتا تازه ای است که با توجه پیشرفت کشور در زمینه IT و نگاه های ویژه به دولت الکترونیک و نفوذ استفاده از سیستم های رایانه ای در صنعت و ایجاد بانک های اطلاعاتی بزرگ توسط ادارات دولتی، بانک ها و بخش خصوصی نیاز به استفاده از آن به طور عمیقی احساس می شود. داده کاوی یعنی کشف دانش و اطلاعات معتبر پنهان در پایگاه های داده. یا به بیان بهتر تجزیه و تحلیل ماشینی داده ها برای پیدا کردن الگوهای مفید و تازه و قابل استناد در پایگاه داده های بزرگ ، داده کاوی نامیده می شود. داده کاوی در پایگاه های داده کوچک نیز بسیار پرکاربرد است و از نتایج و الگوهای تولید شده بوسیله آن در تصمیم گیری های استراتژیک تجاری شرکتهای کوچک نیز می توان بهره های فراوان برد. کاربرد داده کاوی در یک جمله را این گونه می توان بیان کرد : " داده کاوی اطلاعاتی می دهد ، که شما برای گرفتن تصمیم هوشمندانه ای درباره مشکلات سخت شغلتان به آنها نیاز دارید. "

مثالی کلاسیک از کاربرد داده کاوی

اغلب تجارت ها به تصميم گيريهاي استراتژيك و يا اتخاذ خط مشي هاي جديد براي خدمت رساني بهتر به مشتريان نياز دارند. به عنوان مثال فروشگاهها آرايش مغازه خود را براي ايجاد ميل بيشتر به خريد مجدداً طراحي مي كنند. اين مثال به داده هايي در مورد رفتار مصرفي گذشته مشتريان براي تعيين الگوهايي به وسيله داده كاوي، نياز دارند.

برای روشن تر شدن مسئله می توان مثال را اینگونه بیان کرد که در یک فروشگاه زنجیره ای پس از داده کاوی مشخص میشود که درصدی از مشتریان خرید تلویزیون ، میز تلوزیون و گلدان کریستالی را هم در همان روز و بعد از خرید تلوزیون میخرند.مدیر فروشگاه می تواند بلافاصله دستوراتی صادر کند که براساس مدلهای تلوزیون موجود میزهایی و براساس مدل میزها گلدانهای کریستالی برای فروش 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 ملاحضه ای ترقی خواهد کرد.

در واقع ابزار داده كـــاوي، داده را مي گيرد و يك تصوير از واقعيت به شكل مدل مي سازد، اين مدل روابط موجود در داده ها را شرح مي دهد .

برای بهبود بهره وری از یک فروشگاه داده کاوی از داده های انبار داده ، مدل هایی را ارائه میدهد که بيانگر اين هستند كه چه محصولات يا خدماتي، به چه مشتريانی، در چه زماني و از طريق چه كانالي عرضه شود .

بيشتر شركتها، بانكهاي داده اي عظيمي شامل داده هاي بازاريابي، منابع انساني و مالي را دارا هستند. بنابراين، سرمايه گذاري در زمينه انبار داده، يكي از اجزاي حياتي در استراتژي مديريت ارتباط با مشتري است .

رابطه مشتري با زمان تغيير مي كند و چنانچه تجارت و مشتري درباره يكديگر بيشتر بدانند اين رابطه تكامل و رشد مي يابد. چرخه زندگي مشتري چارچوب خوبي براي به كارگيري داده كاوي در مديريت ارتباط با مشتري فراهم مي كند. در بخش ورودي داده كاوي، چرخه زندگي مشتري مي گويد چه

اطلاعاتي در دسترس است و در بخش خروجي آن، چرخه زندگي مي گويد چه چيزي احتمالاً جالب توجه است و چه تصميماتي بايد گرفته شود. داده كاوي مي تواند سودآوري مشتري هاي بالقوه را كه مي توانند به مشتريان بالفعل تبديل شوند، پيش بيني كند و اينكه تا چه مدت به صورت مشتريان وفادار خواهند ماند و چگونه احتمالاً ما را ترك خواهند كرد.

بعضي از مشتريان مرتباً مراجعاتشان را به شركتها براي كسب مزيتهايي كه طي رقابت ميان آنها به وجود مي آيد، تغيير مي دهند. در اين صورت شركتها مي توانند هدفشان را روي مشترياني متمركز كنند كه سودآوري بيشتري دارند.
بنابراين مي توان از طريق داده كاوي ارزش مشتريان را تعيين، رفتار آينده آنها را پيش بيني و تصميمات آگاهانه اي را در اين رابطه اتخاذ كرد.

از کاربرد های داده کاوی می توان به نمونه های زیر اشاره کرد:

.1بانکداری:

- از جالب توجه ترین کاربرد های داده کاوی می توان به کشف پول شویی اشاره کرد .
- تشخیص مشتریان ثابت و همیشگی
- تعیین مشتریان استفاده کننده از یک سرویس خاص

.2بیمه:

- پیش گویی میزان استقبال از بیمه نامه های جدید
- تشخیص کلاهبرداری ها و مشخص کردن رفتار های نا متناسب
- تشخیص نیاز مشتریان و خواسته های آنها
- تشخیص تخلفات پزشکی

واضح است که زمینه استفاده از داده کاوی بی نهایت گسترده است.و دو مثال فوق به خاطر درک راحت تر انتخاب شده اند.
داده کاوی شباهت زیادی به تحلیل های آماری دارد. ولی داده کاوی از جهات زیادی با آمار متفاوت است و مزیت های زیادی نسبت به آمار دارد. جالب ترین تفاوت داده کاوی با تحلیل های آماری این است که در آمار ما فرضیه ای طرح می کنیم و با استفاده از تحلیل های آماری به اثبات یا رد فرضیه می پردازیم اما داده کاوی به فرضیه احتیاجی ندارد.در واقع ابزار داده کاوی فرض می کند که شما خود هم نمی دانید به دنبال چه می گردید. و این نکته ای است که باعث می شود کار آمدی داده کاوی در مواقع بروز مشکل نمایان شود . برای مثال ما در آمار فرض می کنیم که دو گروه فاصله ای باهم ارتباط دارند سپس با استفاداه از ضریب هم بستگی پیرسون مشخص می کنیم که ارتباط وجود دارد یا خیر . ولی داده کاوی بدون توجه به اینکه ما اینگونه فرضی داشته باشیم یا نه با کاوش میان داده ها اگر ارتباطی مخفی معنی داری وجود داشته باشد آن را به اطلاع ما می رساند .تفاوت بعدی آمار و داده کاوی در این است که آمار فقط می تواند از داده های عددی استفاده کند ولی داده کاوی از داده های غیر عددی هم استفاده می کند . تفاوت های دیگری هم میان آمار و داده کاوی وجود دارد که بحث در مورد آنها در حوصله این مقاله نمی گنجد.

اما برای اولین بار در سال ۱۹۵۰ از رایانه برای تحلیل و ذخیره پایگاه داده ها استفاده شد. ولی حجم اطلاعات و میزان رشد آنها به قدری زیاد بوده است که هم اکنون کسی از میزان اطلاعات ذخیره شده در پایگاه داده های سراسر دنیا به صورت دقیق اطلاعی ندارد ولی مطمئنا حجم اطلاعات و مخصوصا

سرعت رشد آنها به قدری زیاد شده که آمار شناسان و تحلیل گران در بررسی و تحلیل پایگاههای داده در زمینه های مختلف ناتوانند .بعضی از پایگاه داده ها به قدری بزرگ و پیچیده شده اند که تحلیل روابط و استخراج اطلاعات مفید پنهان شده در آنها واقعا از ظرفیت ذهنی بشری فراتر رفته است . از زمانی که رشد پایگاه های داده و حجم اطلاعات ، سرعت گرفت و میزان داده ها افزایش یافت ، نیاز به تحلیل ماشینی داده ها و استخراج سریع و دقیق دانش نهفته در آنها احساس شد. شاید بتوان لوول (۱۹۸۳) را اولین شخصي دانست که گزارشي در مورد داده کاوي تحت عنوان « شبیه سازي فعالیت داده کاوي » ارائه نمود .

عمل داده کاوی از یک پایگاه داده به چند مرحله مشخص تقسیم می شود که ما در این مقاله به معرفی و توضیحی مختصر در مورد هر یک از این مراحل اکتفا می کنیم:

.1مرحله اول : تشکیل انبار داده.

با توجه به عنوان ، این مرحله برای تشکیل محیطی پیوسته و یک پارچه جهت انجام مراحل بعدی و داده کاوی در آن، انجام می گیرد.در حالت کلی انبار داده مجموعه پیوسته و طبقه بندی شده است که دائما در حال تغییر بوده و دینامیک است که برای کاوش آماده می شود.

.2مر حله دوم : انتخاب داده ها

در این مرحله برای کم کردن هزینه های عملیات داده کاوی، داده هایی از پایگاه داده انتخاب می شوند که مورد مطالعه هستند و هدف داده کاوی دادن نتایجی در مورد آنهاست.

.3مرحله سوم : تبدیل داده ها.

مشخص است برای انجام عملیات داده کاوی لزوما باید تبدیلات خاصی روی داده ها انجام گیرد ممکن است این تبدیلات خیلی راحت و مختصر مثل تبدیل byte به integer باشد یا خیلی پیچیده و زمان بر و با هزینه های بالا مثل تعریف صفات جدید و یا تبدیل و استخراج داده ها از مقادیر رشته ای و ... باشد.

.4مرحله چهارم : کاوش در داده ها.

در این مرحله است که داده کاوی انجام می شود.در این مرحله با استفاده از تکنیک های داده کاوی داده ها مورد کاوش قرار گرفته ، دانش نهفته در آنها استخراج شده و الگو سازی صورت می گیرد.

.5مر حله پنجم : تفسیر نتیجه.

در این مرحله نتایج و الگو های ارائه شده توسط ابزار داده کاو مورد بررسی قرار گرفته و نتایج مفید معیین می شود.

طرز کار ابزار داده کاو اینگونه است که ابزار به دنبال اثبات این است که وجود چیزی به معنای وجود چیز دیگری است و سعی می کند در درجه اول از توالی ارتباطات برای کشف یک الگو بهره بگیرد و در نهایت اطلاعات بدست آمده را دسته بندی کند تا به الگوی خاصی برسد که بتواند آن را براساس فاکتورهای داخی به مخاطبش ارائه دهد.

همچنین در داده کاوی از الگوریتم های ژنتیک و شبکه های عصبی هم استفاده می شود.شبکه های عصبی به علت کار آمدی در حل مسائل پیچیده و بزرگ مورد استفاده اند و کاربرد الگوریتم های ژنتیک در داده کاوی برای جستجو و ساختن یک مدل بهینه در میان مدل های بدست آمده است ، به این گونه که مدل های اولیه روی کرومزوم هایی قرار می گیرند و با رقابت بر سر انتقال صفات به نسل بعد ، بهترین مدل و لایق ترین آنها به کاربر ارائه می شوند.

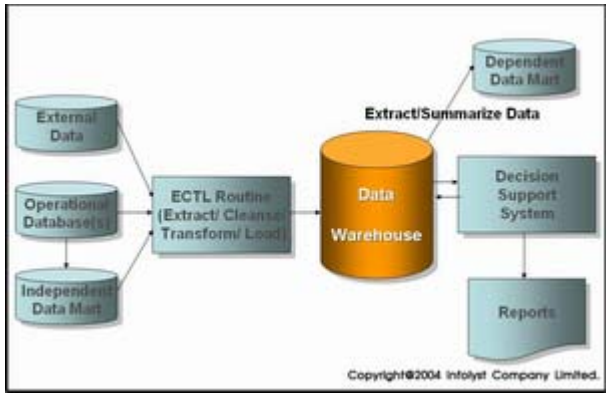
داده کاوی امروز گسترش زیادی یافته است به طوری که اکثر نرم افرار های پایگاه داده ای مثل SQL Serverو ORACLE نیز شامل ابزارهایی داده کاوی شده اند ولی هنوز نرم افزار های تخصصی داده کاوی همچون Intelligent Miner , Darwin , Mine Set, Knowledge Studio, Data Mind از مهمترین ابزار های داده کاوی اند.

آشنایی با مفاهیم انباره های داده(DatawareHouse)

کاوی داده / انباره داری / تحلیل زنده

مقدمه

آوری و فن (DatawareHouse) **داده** مفاهیم انباره های با آشنایی خواننده **مقاله** هدف از ارایه این برای ساخت و بهره برداری از برنامه پایه گذار فن آوری لازم **داده** های مرتبط می باشد. انباره های تخیلی با آنها آشنا شده ایم. برنامه هایی که های هوشمندی هستند که بعضا در برخی فیلم های .نموده و پس از تجزیه و تحلیل با آنها به مشاوره می پردازند اطلاعات را برای صاحبان خود جمع آوری



چیست؟ OLTP

در دهه فقط در سال ۲۰۰۰ میزان ظرفیت نصب شده جهت ذخیره سازی اطلاعات از کل ظرفیت موجود .۱۹۹۰ بیشتر بوده است

تقریبا حجم کل اطلاعات در کامپیوترها هر ۵ هاست. در حال حاضر **داده** حیات بازرگانی نوین مبتنی بر ایجاد برنامه های چند رسانه ای و بانکهای اطلاعاتی پیش سال دو برابر می شود و با توجه به سرعت .اطلاعات به دو برابر در سال برسد بینی می شود که شتاب رشد
جدیدی هستند که امور خود را توسط کامپیوترها تولید کنندگان این اطلاعات موسسات و شرکت های جمع آوری نموده و به مصرف می ها را **داده** مکانیزه ای که هدایت می کنند. سیستم های تولید .ها هستند **داده** تولید کنندگان واقعی نامیده می شوند. این سیستم ها OLTPرسانند سیستم های

:برنامه های کاربردی خادم و مخدوم بدو دسته تقسیم می شوند
(DSS) پشتیبانی تصمیم گیری سیستم های "
اطلاعات (OnLine) سیستم های پردازش زنده "

قبل از آنکه به .این دو دسته هر یک راه های کاملا متفاوتی را جهت حل مسایل تجاری ارایه می کنند .تفاونهای این دو را بشناسیم پی ببریم لازم است **داده** ارزش انباره های

رزرواسیون، در کلیه خدمات بازرگانی دیده می شوند از جملع سیستم های OLTP سیستم های به زمان پاسخی بین ۱ تا دستگاه های فروش ، کنترل انبار، سهام و فروش و ... . این سیستم ها غالبا در ساعات مختلف روز ، هفته و ماه می تواند ۳ ثانیه در ۱۰۰ در صد اوقات نیاز دارند. تعداد کاربران آنها همان زمان پاسخ قبلی نیاز دارند. در این گونه سیستم ها بشدت متغیر باشد و درتمامی این اوقات به به خادمین تعاملی (Database Servers)با بانکهای اطلاعاتی معمولا مخدومین بجای ارتباط متصل می شوند. البته این گونه ارتباط لازمه دستیابی به سرعت مورد نیاز (Transaction Servers) .است (Clients) مخدومین

قادرند تعامل را تقسیم می گردد. خادمین عادی (Heavy) و قوی (Lite) خود نیز به دو نوع عادی OLTP به اجرا بگذارند و خادمین قوی (StoredProcedures) در غالب پردازش های ثبت شده در بانک اطلاعاتی سربار ،برای دستیابی به سرعت OLTP میکنند.در برای اجرای دستورات استفاده (TP Monitor) از انتقال یک دستور ارتباطی شبکه ها در حداقل ممکن نگاه داشته می شود و غالبا ارتباطات در حد .سیکویل هستند SQL))
بسرعت پایگاه های اطلاعاتی بزرگی یا با جمع آوری اطلاعات امروزه حتی کوچکترین تجارتها هم قادرند چه رسد به وب سرور ها که می توانند ظرف مدت بسیار کوتاهی چندین صندوق های فروش ایجاد کنند .جمع آوری نمایند گیگا بایت اطلاعات

امروزه هر کسی بسادگی زمانی برای هر کار مکانیزه ای نیاز به میلیونها پول و ده ها متخصص بود . اما از امکانات رایانه ای بهره مند گردد. با خرید چند کامپیوتر شخصی و استخدام یک برنامه نویس می تواند ها برای همگان **داده** ایجاد پایگاه های خصوصی از بعبارت دیگر دسترسی به خدمات رایانه ای برای .شده است آسانتر

جمع آوری می شود مستقیما مورد استفاده OLTP توسط سیستم های هایی که **داده** در مجموع دانند چگونه نیاز ها چیستند و همچنین می **داده** قرار دارد. آنها دقیقا می دانند این افراد ایجاد کننده آن .کنند های اطلاعاتی لحظه ای خود را که بطور روزمره بوجود می آید حل

داشته باشد به این اطلاعات نیاز OLTP سوالی که مطرح است اینست که اگر کسی خارج از مجموعه بایستی آنرا پیدا کرد و چگونه به ای موجود است؟ کجا **داده** چه باید کرد. این افراد از کجا می دانند چه است ؟ چه معنایی دارد؟ آخرین چیزی که (Format) ها به چه شکلی **داده** آن دسترسی پیدا کنند؟ اجازه دهند دیگران به اطلاعات گرانبهای آنان دسترسی به آن رضایت خواهند داد آنست که OLTP افراد دانند چه می خواهند، درخواستهای سیکویل زمانگیری را بر روی داشته باشند. کسانی که حتی نمی .آورد ها را پایین می **داده** می کنند که سرعت و قابلیت سیستم تولید کننده بانکهای اطلاعاتی اجرا

مشابه خود در خود می خواستند با همکاران MIS در گذشته افراد بیرون از سیستم ، از همکاران استخراج نمایند. اما امروزه حتی سیستم مربوطه تعامل داشته و نهایتا اطلاعات مورد نظر را از سیستم سازمان موجود است. اطلاعات بشدت توزیع خود هم بدرستی نمی داند چه اطلاعاتی در MISمجموعه .بخشی از اطلاعات سازمان وجود دارد شده و پراکنده است و تقریبا روی هر کامپیوتری

افراد یکی از ویژگیهای کامپیوتر های شخصی و همچنین معماری خادم/مخدوم موجب شده است که ترجیح می دهند اطلاعات اکثرا به اطلاعات سازمانی و کاربرد اطلاعات در سازمان علاقه ای نداشته و سازمان و شخصی (یا واحد های را تحت مالکیت شخصی اداره کنند به این ترتیب بین اطلاعات استخراج های سیستمی و اطلاعات **داده** متشکله) شکاف وجود خواهد داشت. از طرف دیگر بین می کنند افرادی شده نیز شکاف دیگری مشاهده می شود. کسانی که از بیرون به این اطلاعات نگاه بتوانند تصمیمات بهتری ها هستند بطوریکه **داده** هستند که بدنبال یافتن طرحها، روالها و تمایلات در تجارت دیگران است و خیلی زود همگان بگیرند. تنیدن حصار بدور اطلاعات بمعنی تنیدن حصار در برابر .بازنده جنگ این حصارها خواهند بود

.چگونه اطلاعات را در اختیار داریم اگر بدیگران اجازه دسترسی به آنرا بدهیم
:اند شوند و از آن جمله **داده** سولات زیادی مطرح هستند که بایستی پاسخ
نمی کند؟ چگونه مطمین شویم که عملکرد بیرونی ها (غریبه ها) عملکرد سیستم ما را کند
چه اطلاعاتی را بایستی در اختیار بیرونی ها قرار دهیم؟
است؟ (**داده** درونی و شخصی (فقط مربوط به سیستم تولید کننده چه اطلاعاتی
اطلاعات به اشتراک گذاشته شده است؟ چه کسی مالک
چه کسی این اطلاعات را بروز میکند؟
کنیم؟ بایستی بگذاریم دسترسی به اطلاعات مستقیم باشد یا آنرا در بانک دیگری کپی آیا
اطلاعات استخراج شده چگونه نگهداری شده و چگونه بروز می شود؟

بشناسیم و تفاوتهای برای پاسخ به سوالات فوق بایستی نیاز های استفاده کنندگان از این اطلاعات را .را درک کنیم OLTP میان سیستمهای پشتیبان تصمیم گیری و

میکنند؟ ها استفاده داده چه کسانی از این

اطلاعات هستند(کسانی بیایید نامی برای این دسته از افراد انتخاب کنیم. این افراد مصرف کنندگان شکارچی اطلاعات می گذاریم چون هستند که تصمیمات استراتژیک می گیرند) فعلا نام این افراد را

نیازمند اطلاعات است. البته بازرگانان و دسترسی دارد و PCاین نام معرف هر کسیست که به یک .هستند صنعتگران اولین دسته از این افراد

سیستم پشتیبانی تصمیم گیری چیست؟

ها، تولید گزارش های **داده** ها ، یافتن ارتباط بین **داده** یک سیستم کارآمد، ابزاریست برای تحلیل پاسخ به قابلیت اطلاعات در انواع ممکن، ها، راهکار های نمایش **داده** کارآمد، دسترسی منعطف به . ها به صفحات گسترده **داده** سوالات اگر ... چه ، چاپ اطلاعات،انتقال
بیشتری در زمان پاسخگویی برخوردار **داده**، این ابزارها از انعطاف در مقایسه با سیستم های تولید است و قابلیت دسترسی همزمان کاربران به آن هستند. معمولا کنترل یکپارچگی در آنها رعایت نشده بروز رسانی اطلاعات غالبا بمعنی پردازش روی تمامی غالبا محدود است. جستجوی اطلاعات و یا برای غیر برنامه نویسان تهیه شده و بیشتر فعالیت ها در آن از طریق اطلاعات خواهد بود. این برنامه ها .انجام می شود (Point and Click) کن نشان بده و کلیک

سیستم های اطلاعات مدیران اجرایی (Executive Information Systems)

همچنین به یک زمینه .قوی تر، ساده تر و کار آمدتر هستند DSS این دسته از برنامه ها از ابزارهای بتدریج کم رنگ شده EIS و DSS تجاری خاص نزدیکتر و طبیعتا گرانتر هم هستند. البته اختلاف بین خود را (Enterprise) و در سطح سازمان **داده** بتازگی دامنه عمل خو.د را گسترش EIS است. ابزارهای استفاده می کنند مطرح کرده اند بطوریکه مدیران و تحلیلگران نیز از این ابزار ها

(MDA)ابزارهای یا OnLine Analytical Processing (OLAP) بطور خلاصه ابزارهای DSS/ESS ابزارهای. (Data **کاوی** **داده** آنها ابزارهای نامیده می شوند و در لایه های بالاتر به Multidimensional Analysis .گفته می شود (Intelligent Agent) کارآگاهان شخصی و (Mining

مقایسه سیستم های OLTP و DSS

:را می بینیم OLTP و DSS در جدول زیر تفاوت های دو نوع سیستم

DSS نیاز بانک اطلاعاتی OLTP قابلیت نیاز بانک اطلاعاتی
می کند کارکنان سیستم تولید کننده اطلاعات شکارچی اطلاعات چه کسی از آن استفاده
مقدار فعلی اطلاعات نیاز دارد و گزارش ها قابل باز سازی نیستند به اطلاعات ارزش زمانی اطلاعات به
هستند نیاز دارد . اطلاعات هر از گاه به وقت می شوند. گزارش ها قابل بازسازی پایدار
از گاه تعداد دسترسی ها به اطلاعات پیوسته در طول روز کاری با نقاط پیک کاری هر
تبدیلی صورت نگرفته در چندین لایه تبدیل صورت گرفته است. خام است. استخراج و **داده** **داده** شکل
ها انجام شده **داده** استخراج و فشرده سازی
از چندین محل داخلی و خارجی ها از یک برنامه **داده** آوری جمع
شود خیر از برنامه های توسط یک برنامه تولید می **داده** مشخص است بلی بیشتر **داده** آیا محل تولید
مختلف و بانک های اطلاعات و وب می آید
ها پیوسته و در یک نگارش هستند بلی هر مجموعه **داده** .بندی شده هستند خیر آیا اطلاعات نگارش
است دارای تاریخ برداشت **داده** از
چندین کاربر اطلاعات را به وقت می کنند بیشتر اوقات یک کاربر **داده** نوع دسترسی به
مقدار کنونی مدام در حال تغییر است فقط خواندنیست قابل به وقت رسانی است **داده** آیا
ندارد. فقط از طریق برنامه ها ممکن است. منعطف از طریق یک تولید انعطاف در دسترسی انعطاف
OLAP کننده درخواست و
نسبتا کند راندمان سرعت پاسخ بالا مورد نیاز است. فعالیت ها همگی مکانیزه و سریع
زیادی کار کشف و تحقیق و نیازهای اطلاعاتی بخوبی فهمیده شده اند ناپایدار و نسبی. به مقدار
.جستجوی موضوعی نیاز است
ها ممکن است از هر جایی بیایند **داده** که در بانک موجود است دامنه اطلاعات محدود. آن چیزی
هزاران و میلیونها رکورد / رکورد های پردازش شده کمتر از ۱۰ رکورد صدها

داده انباره (DatawareHouse)

اطلاعات برای مصرف سیستم های (Repository) یعنی انباره **داده** در محیط خادم/مخدوم انباره
از اطلاعات است که قادر است اطلاعات بک مخزن فعال و هوشمند **داده** پشتیبانی تصمیم گیری.انباره

و نهایتا پخش نماید و در صورت لزوم نیز سیاست را از محیط های گوناگون جمع آوری و مدیریت کرده نماید های تجاری را روی آنها اجرا.

عناصر انباره داری

انباره یک محل است و انباره داری یک فرآیند.

این فرآیند از عناصر زیر تشکیل شده است :

۱. روی بانک های مختلف مدیریت انتشار اطلاعات انباره که وظیفه نسخه برداری و توزیع اطلاعات را بر اطلاعاتی را که بایستی کپی (آنگونه که شکارچی اطلاعات تعریف می کند) به عهده دارد. شکارچی تبدیلات لازم روی اطلاعات را تعریف می کند. شود، مبدا و مقصد اطلاعات، تعداد بوقت رسانی ها و کامل آخرین وضعیت اطلاعات و اصطلاح بوقت رسانی بمفهوم کپی (Refresh) اصطلاح تازه سازی بکار گرفته شده اند. همه کارها می تواند بصورت خودکار و یا بمفهوم اعمال آخرین تغییرات (Update) اطلاعات ممکن است از بانکهای رابطه ای و غیر رابطه ای تهیه شود. توجه کنید که .دستی انجام پذیرد اطلاعات خارجی قبل از ورود به سیستم، تبدیل شده و پاک سازی می شوند کلیه.

۲. نمودن یک بانک اطلاع رسانی یک بانک اطلاعاتی رابطه ایست که وظیفه سازماندهی و ذخیره حاصله از منابع مختلف و نسخه از اطلاعات و همچنین تبدیلات و جمع بندی و افزودن ارزش به اطلاعات در مورد اطلاعات) نیز به عهده این بانک با فرمت های مورد نظر بعهده دارد. نگهداری فراداده (اطلاعات ایندکس ها و غیره را بیان می کنند و فراداده های است . فراداده های سیستمی روابط بین جداول و را برای یک شکارچی اطلاعات روشن می سازند ارزش اطلاعات (semantic) محتوایی.

۳. راهنمای تجاری و یک ترکیبی از یک راهنمای فنی و (Informational Directory) راهنمای اطلاعات دانستن محل وجود اطلاعات پویشگر اطلاعات است. هدف اصلی این راهنما کمک به شکارچی برای ،شکل آن و روش دسترسی به آن است

۴. از فروشندگان انجام می گیرد. بسیاری SQL از طریق انواع دستورات DSS/EIS پشتیبانی ابزارهای و سایرین انواع دیگر پروتکل ها را سرویس می دهند ODBC پروتکل.

)(DataMarts) داده سلسله مراتب انباره ها)غرفه های

داده دپارتمانی و غرفه های **داده** هستند. در عمل غرفه های **داده** انواع کوچکتری از انباره های ابتدا برنامه ریزی نمی شوند بلکه ابتدایا بوجود آمده و در صورت موفقیت تکثیر شده از (mobile) همراه ها تشکیل در نهایت مدیر بانک اطلاعاتی سازمان ممکن است بتواند یک فدراسیون آزاد ار این غرفه و نماید را پایه گذاری **داده** دهد و نهایتا یک انباره.

(DataMining) **کاوی** **داده** و تا (OLAP) تا تحلیل زنده (Queries) از خواسته ها DSS/EIS ابزارهای

ابزارهای گزارش گیری

خواسته پردازها بما اجازه ساختن یک دستور سیکویل را می دهند بدون آنکه و **داده** ابزارهای تحلیل برنامه ای بنویسیم یا سیکویل یاد بگیریم. با چند نشانه و کلیک عبارت های سیکویل مجبور باشیم شود. مناسب برای گرد آوری اطلاعات و نمایش آن بشکل یک گراف / جدول و یا گزارش آماده می را می دهند و به این ابزارهای برجسته تر در این زمینه امکان کنترل میزان نتایج برگشته از یک خواسته را برگردانند گرفت. در سال ۱۹۹۸ ترتیب می توان جلوی درخواستهایی را که ممکن است میلیونها رکورد Microsoft Access, Oracle Reports, است که بیش از ۱۵۰ نوع از این ابزارها در بازار وجود داشته از آن جمله اند Business Objects.

و اطلاعات چند بعدی OLAP

که می توانید آنرا در جهات مختلف نگاه کنید ها **داده** مثل یک مکعب روبیک از OLAP به ساختار کنید بچرخانید تا بتوانید سناریو های "قبلا چه شده" و "چه می شد اگر ..." را بررسی.

اطلاعاتی دو بعدی (و یا بانکهای خاص ها را توسط بانکهای **داده** این ابزارها دیدگاههای چند بعدی از قدرت OLAP ها در **داده** توان دسترسی چند بعدی به .چند بعدی) تولید کرده و در اختیارما می گذارند .کردن خواسته های پیچیده تر را بما می دهد فرموله

متعارف فقط یک صفحه گسترده با چند محور است (در صفحات گسترده OLAP برای سادگی فرض کنید داریم) در این صورت مثلا می ...,و عمودی با ایندکس های ۱و۲و۳ ... A, B, C, دو محور افقی با اختصار فروش، تاریخ، مشتری، فروشگاه، قیمت و توانیم اطلاعات فروش یک سازمان را از دیدگاه های منطقه میزان فروش به ازای یک محصول و فروشگاه در یک میزان فروش بررسی کنیم. و پاسخ سولاتی نظیر .ماه مشخص را خواهیم داشت

های اطلاعاتی رابطه ای ها را در مقایسه با بانک **داده** طریقه نمایش دادن OLAP مدل چند بعدی اطلاعاتی رابطه ای سرویس فوق را ارایه با ایجاد یک لایه محافظ روی یک بانک ROLAP.تسهیل می کند ذخیره سازی و محاسبه اطلاعات چند بعدی برای فقط راهی برای OLAP میدهد. از دیدگاه فنی پیش روی چندین محور جمع می ها را از **داده**، OLAP خادم پاسخوگویی به سناریوهای کاربر است. یک ها **داده** OLAP بایستی پاک سازی شوند . غالبا OLAP زند. توجه کنید که اطلاعات قبل از وارد شدن به .استخراج می کند **داده** را از یک انباره

:را به چند دسته تقسیم می کنند OLAP ابزارهای

:رو میزی OLAP

سازند ابزارهای ساده و مستقل که روی کامپیوتر های شخصی نصب شده و مکعب های کوچکی می صفحات گسترده ای و آنها را نیز بر روی سیستم به شکل فایل ذخیره می کنند. بیشتر این ابزارها با به استفاده از این دسته از کار می کنند.به این ترتیب کسانی که در سفر هستند قادر Excel نظیر (کردن این محصولات است در حال جایگزین Web OLAP محصولات هستند.(در حال حاضر

چند بعدی MOLAP

اطلاعاتی خاصی را بجای ذخیره کردن اطلاعات در رکورد های کلید دار، این دسته از ابزارهای بانکهای **داده** مرتب شده بر اساس ابعاد ها را به شکل آرایه های **داده** برای خود طراحی کرده اند بطوریکه در حال حاضر نیز دو استاندارد برای این تیپ ابزار وجود دارد. سرعت این (HyperCubes) ذخیره می کنند .ابزار بالا ولی سایز بانک اطلاعاتی آن نسبتا کوچک است

(ROLAP) رابطه ای OLAP

می کنند. بطوریکه این ابزار ها با ایجاد یک بستر روی بانکهای رابطه ای اطلاعات را ذخیره و بازیابی بر همین Red Brick, MicreoStrategy اساس بهینه سازی برخی بانکهای اطلاعاتی رابطه ای مانند .اساس استوار است
.توجه می باشد اندازه بانک اطلاعاتی این ابزار قابل

Hybrid OLAP (HOLAP)

است (MOLAP طرح شده در) MDBMS و ROLAP ترکیبی از hybrid در اینجا منظور از .می باشد ROLAP ابزار دارای بانک اطلاعاتی بزرگ و رادمان بالاتر نسبت به

OLAP استانداردهای

و از طرف دیگر MD-API با استاندارد OLAP با دو استاندارد مواجه است، از یک طرف گروه OLAP جامعه و دومی از حمایت Oracle اولی از حمایت (Tensor) OLE DB for OLAP با استاندارد Microsoft برنامه های آنان را در ابعاد فروش MS-SQLV فروشندگان کوچکتری برخوردار است که امیدوارند فروش .ویندوز مطرح کند

(Data Mining)کاوی داده

طلای کوچکی را که در های ما می توانند تکه **داده** با جستجوی حجم عظیم **کاوی** **داده** ابزارهای .گوشه ای پنهان شده بیابند

های یک واحد **داده** بازگشت هزینه صرف شده در این ابزارها غالبا بسیار سریع است. مثلا در بررسی فروشگاه متوجه شدند که میزان سرقت حین فروش از باتریها و فیلمها و قلم های با قیمت ار یک جابجاکردن متوسط ماهانه حدود ۶۰۰۰۰ دلار برای فروشگاه هزینه داشته است که به این ترتیب با .جویی بدنبال داشته است اقلام و قرار دادن در قسمتهای با دید بهتر سالانه حدود ۷۰۰۰۰۰ دلار صرفه

گردد که ممکن است از دید ما ها می **داده** بندی هایی در بدنبال طرحها و گروه **کاوی** **داده** ابزارهای که استفاده OLAP گیرد. بر خلاف ابزارهای پنهان مانده باشد.. ابزار تقریبا از کاربر هیچ کمکی نمی استفاده کننده را این ابزار است که **کاوی** **داده** هستند در کنندگان راهنما و سازمان دهنده اطلاعات که چه می خواهید. بیشتر این هدایت می کند. ابزار فرض می کند که شما خود نیز دقیقا نمی دانید :ابزار ها از روش های جستجوی زیر استفاده میکنند
۱- است که وجود ارتباطات که اصطلاحا تحلیل سبد بازار خوانده می شود. ابزار بدنبال اثبات این موضوع تعطیلات تابستانی در استرالیا چیزی بمعنی وجود چیز دیگریست. مثلا بیشتر خریداران لوازم غواصی به .خریدار کا لای دیگری نیز هست می روند. یا مصرف کننده یک کالای مشخص مصرف کرده

۲- قیمت طلا ۱۰ درصد ارتباطات متوالی ابزار بدنبال روابط متوالی بین موضوعات می گردد مثلا وقتی .بالا می رود یک هفته بعد قیمت سهام ۱۵ درصد پایین می آید

۳- درصد رای دسته بندی بدنبال دسته بندی و طبقه بندی سطح بالای اطلاعات هستند. مثلا ۷۰ دارند بین ۴۰ تا ۵۰ سال دهندگانی که تصمیم نگرفته اند به که رای دهند درآمدی بالای ۶۰۰۰۰ دلار .اقامت دارند X سن دارند و در منطقه

می رسید افراد اگر اطلاعات جدول زیر در یک گراف دو بعدی به تصویر در آید متوجه می شویم که بنظر بین ۲۳ تا ۲۹ به مکزیک و بین ۳۰ تا ۵۱ به کانادا سفر می کنند

سن مشتری کشوری که به آن سفر کرده
مکزیک ۲۳
کانادا ۴۵
کانادا ۳۲
کانادا ۴۷
کانادا ۴۶
کانادا ۳۴
کانادا ۵۱
مکزیک ۲۸
کانادا ۴۹
مکزیک ۲۹
مکزیک ۲۶
کانادا ۳۱

نمی کنند یک نکته جالب دیگر که بسادگی قابل دیدن نیست آنستکه افراد بین ۳۵ تا ۴۴ اصلا سفر آنها که بین ۴۵ تا ۵۱ سال بعبارت دیگر دو دسته آدم به کانادا سفر می کنند آنها که بین ۳۰ تا ۳۴ و بعدی بسادگی قابل رویت است . چنانچه سن دارند. گروه بندی در این مجموعه اطلاعات کوچک و دو سادگی گذشته نخواهد بود. گفتنی است تعداد نمونه ابعاد اطلاعات و حجم آن افزایش یابد موضوع به **کاوی** **داده** مقادیری که هر یک از ستونها می گیرند در سرعت پردازش ها، تعداد ستونهای اطلاعاتی و پردازش ۱۰۰۰۰۰۰ نمونه با ۲۰۰ ستون اطلاعاتی که هر یک می توانند ۲۵ موثر هستند. مثلا برای . خود بگیرند به حدود ۲ ساعت وقت نیازاست مقدار مختلف به
کاربرد یافته اند.از جمله محققین بهداشت برای کشف میزان موفقیت این ابزار ها در زمینه های مختلف برای ارزیابی اعتبار مشتریان، بورس بازان برای تشخیص جابجایی قیمتهای سهام و جراحیها . بانکها طرحهای تجاری ، شرکتهای بیمه برای تشخیص ریسک مشتریان و رفتارهایشان و هتل ها تشخیص **داده** تشخیص مشتریان بازگشتی خود از آن استفاده میکنند. همانطوریکه بنظر می آید ابزارهای برای رده بالاتر هستند که استفاده های قابل توجهی برای آنها در صنعت قابل از مجمیعه ابزارهای یک **کاوی** .تصور است

: برخی از انواع تجاری این ابزار عبارتند از
Intelligent Miner, Darwin, MineSet, KnowledgeStudio, DataMind, Clementine
, ...

کارآگاهان شخصی

اطلاعات مورد نیاز را **داده** این مامورین برنامه های قابل حملی هستند که با اتصال به انباره های قوانین تعریف استخراج کرده و به کارفرمایان خود اطلاع می دهند. در حال حاضر این ابزارها بر اساس مشاهده تغییر پیغام شده از طرف کارفرمای خود به جستجوی تغییرات در اطلاعات رفته و در صورت .مناسب را می دهند

: هنوز کار های زیادی در این قسمت بایستی صورت پذیرد که از آن جمله اند
ابزار (هوشمند شدن)، درک علایق کارفرما و جستجو دربانکهای ها بر اساس دانش درون **داده** درک .اعلام تغییرات به کارفرما اطلاعاتی مختلف برای
کاربر و یا برنامه های همسر یابی با توجه به برنامه های کاریابی روی اینترنت با توجه به رزومه .... و سایر شرکتها اعلام می کنند و Microsoft سایتهای مشخصات. برنامه هایی که تغییرات را در .هستند از این نوع برنامه ها مثالهای ساده ای

کپی رایت و منابع

: مقالات مرتبط
.داشته باشد، موجود نیست **مقاله** ارتباط زیادی با این ای که **مقاله** در حال حاضر متاسفانه

منتشر کامپیوتر ب.ظ و درباره موضوعات ۳:۴۶ این مطلب در تاریخ سه شنبه ۹ آبان ۱۳۸۵ در ساعت نظر میتوانید دنبال کنید. شما RSS 2.0 شما میتوانید هر پاسخی به این مطلب را توسط .شده است تالارهای گفتگو لطفاً سوالات و درخواست های خود را در .ارسال کنید دنبالک ، یا از سایت خودبدهید .نمائید مطرح

.شده است داده (DatawareHouse)" داده یک نظر درباره "آشنایی با مفاهیم انباره های

1. گفت *محسن بختیاری*
ساعت ۱۲:۵۹ ب.ظ چهارشنبه ۲۶ اردیبهشت ۱۳۸۶

تشکر ای می خواستم .با **مقاله** در مورد فایلهاس سیستمی

استاد راهنما : مهندس پیمان عمادی

معرفی داده کاوی

داده کاوی [1] یا کشف دانش در پایگاه داده ها [2] (**KDD**) علم نسبتا تازه ای است که با توجه پیشرفت کشور در زمینه IT و نگاه های ویژه به دولت الکترونیک و نفوذ استفاده از سیستم های رایانه ای در صنعت و ایجاد بانک های اطلاعاتی بزرگ توسط ادارات دولتی، بانک ها و بخش خصوصی نیاز به استفاده از آن به طور عمیقی احساس می شود. داده کاوی یعنی کشف دانش و اطلاعات معتبر پنهان در پایگاه های داده. یا به بیان بهتر تجزیه و تحلیل ماشینی داده ها برای پیدا کردن الگوهای مفید و تازه و قابل استناد در پایگاه داده های بزرگ ، داده کاوی نامیده می شود. داده کاوی در پایگاه های داده کوچک نیز بسیار پرکاربرد است و از نتایج و الگوهای تولید شده بوسیله آن در تصمیم گیری های استراتژیک تجاری شرکتهای کوچک نیز می توان بهره های فراوان برد. کاربرد داده کاوی در یک جمله را این گونه می توان بیان کرد : " داده کاوی اطلاعاتی می دهد ، که شما برای گرفتن تصمیم هوشمندانه ای درباره مشکلات سخت شغلتان به آنها نیاز دارید" [3] .

مثالی کلاسیک از کاربرد داده کاوی

اغلب تجارت ها به تصميم گيريهاي استراتژيك و يا اتخاذ خط مشي هاي جديد براي خدمت رساني بهتر به مشتريان نياز دارند. به عنوان مثال فروشگاهها آرايش مغازه خود را براي ايجاد ميل بيشتر به خريد مجدداً طراحي مي كنند. اين مثال به داده هايي در مورد رفتار مصرفي گذشته مشتريان براي تعيين الگوهايي به وسيله داده كاوي، نياز دارند.

برای روشن تر شدن مسئله می توان مثال را اینگونه بیان کرد که در یک فروشگاه زنجیره ای پس از داده کاوی مشخص میشود که درصدی از مشتریان خرید تلویزیون ، میز تلوزیون و گلدان کریستالی را هم در همان روز و بعد از خرید تلوزیون میخرند.مدیر فروشگاه می تواند بلافاصله دستوراتی صادر کند که براساس مدلهای تلوزیون موجود میزهایی و براساس مدل میزها گلدانهای کریستالی برای فروش 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 ملاحضه ای ترقی خواهد کرد.

در واقع ابزار داده كـــاوي، داده را مي گيرد و يك تصوير از واقعيت به شكل مدل مي سازد، اين مدل روابط موجود در داده ها را شرح مي دهد.

برای بهبود بهره وری از یک فروشگاه داده کاوی از داده های انبار داده ، مدل هایی را ارائه میدهد که بیانگر این هستند که چه محصولات يا خدماتي، به چه مشترياني، در چه زماني و از طريق چه كانالي عرضه شود.

بيشتر شركتها، بانكهاي داده اي عظيمي شامل داده هاي بازاريابي، منابع انساني و مالي را دارا هستند. بنابراين، سرمايه گذاري در زمينه انبار داده، يكي از اجزاي حياتي در استراتژي مديريت ارتباط با مشتري است.

رابطه مشتري با زمان تغيير مي كند و چنانچه تجارت و مشتري درباره يكديگر بيشتر بدانند اين رابطه تكامل و رشد مي يابد. چرخه زندگي مشتري چارچوب خوبي براي به كارگيري داده كاوي در مديريت ارتباط با مشتري فراهم مي كند. در بخش ورودي داده كاوي، چرخه زندگي مشتري مي گويد چه اطلاعاتي در دسترس است و در بخش خروجي آن، چرخه زندگي مي گويد چه چيزي احتمالاً جالب توجه است و چه تصميماتي بايد گرفته شود. داده كاوي مي تواند سودآوري مشتري هاي بالقوه را كه مي توانند به مشتريان بالفعل تبديل شوند، پيش بيني كند و اينكه تا چه مدت به صورت مشتريان وفادار خواهند ماند و چگونه احتمالاً ما را ترك خواهند كرد.

بعضي از مشتريان مرتباً مراجعاتشان را به شركتها براي كسب مزيتهايي كه طي رقابت ميان آنها به وجود مي آيد، تغيير مي دهند. در اين صورت شركتها مي توانند هدفشان را روي مشترياني متمركز كنند كه سودآوري بيشتري دارند.

بنابراين مي توان از طريق داده كاوي ارزش مشتريان را تعيين، رفتار آينده آنها را پيش بيني و تصميمات آگاهانه اي را در اين رابطه اتخاذ كرد.

از کاربرد های داده کاوی می توان به نمونه های زیر اشاره کرد :

۱. بانکداری :

- از جالب توجه ترین کاربرد های داده کاوی می توان به کشف پول شویی اشاره کرد.
- تشخیص مشتریان ثابت و همیشگی
- تعیین مشتریان استفاده کننده از یک سرویس خاص

۲. بیمه :

- پیش گویی میزان استقبال از بیمه نامه های جدید
- تشخیص کلاهبرداری ها و مشخص کردن رفتار های نا متناسب
- تشخیص نیاز مشتریان و خواسته های آنها
- تشخیص تخلفات پزشکی

واضح است که زمینه استفاده از داده کاوی بی نهایت گسترده است.و دو مثال فوق به خاطر درک راحت تر انتخاب شده اند.

داده کاوی شباهت زیادی به تحلیل های آماری دارد. ولی داده کاوی از جهات زیادی با آمار متفاوت است و مزیت های زیادی نسبت به آمار دارد. جالب ترین تفاوت داده کاوی با تحلیل های آماری این است که در آمار ما فرضیه ای طرح می کنیم و با استفاده از تحلیل های آماری به اثبات یا رد فرضیه می پردازیم اما داده کاوی به فرضیه احتیاجی ندارد.در واقع ابزار داده کاوی فرض می کند که شما خود هم نمی دانید به دنبال چه می گردید. و این نکته ای است که باعث می شود کار آمدی داده کاوی در مواقع بروز مشکل نمایان شود . برای مثال ما در آمار فرض می کنیم که دو گروه فاصله ای باهم ارتباط دارند سپس با استفاداه از ضریب هم بستگی پیرسون مشخص می کنیم که ارتباط وجود دارد یا خیر . ولی داده کاوی بدون توجه به اینکه ما اینگونه فرضی داشته باشیم یا نه با کاوش میان داده ها اگر ارتباطی مخفی معنی داری وجود داشته باشد آن را به اطلاع ما می رساند .تفاوت بعدی آمار و داده کاوی در این است که آمار فقط می تواند از داده های عددی استفاده کند ولی داده کاوی از داده های غیر عددی هم استفاده می کند . تفاوت های دیگری هم میان آمار و داده کاوی وجود دارد که بحث در مورد آنها در حوصله این مقاله نمی گنجد.

اما برای اولین بار در سال ۱۹۵۰ از رایانه برای تحلیل و ذخیره پایگاه داده ها استفاده شد. ولی حجم اطلاعات و میزان رشد آنها به قدری زیاد بوده است که هم اکنون کسی از میزان اطلاعات ذخیره شده در پایگاه داده های سراسر دنیا به صورت دقیق اطلاعی ندارد ولی مطمئنا حجم اطلاعات و مخصوصا سرعت رشد آنها به قدری زیاد شده که آمار شناسان و تحلیل گران در بررسی و تحلیل پایگاههای داده در زمینه های مختلف ناتوانند. بعضی از پایگاه داده ها به قدری بزرگ و پیچیده شده اند که تحلیل روابط و استخراج اطلاعات مفید پنهان شده در آنها واقعا از ظرفیت ذهنی بشری فراتر رفته است . از زمانی که رشد پایگاه های داده و حجم اطلاعات ، سرعت گرفت و میزان داده ها افزایش یافت ، نیاز به تحلیل ماشینی داده ها و استخراج سریع و دقیق دانش نهفته در آنها احساس شد. شاید بتوان لوول (۱۹۸۳) را اولین شخصي دانست که گزارشي در مورد داده کاوي تحت عنوان « شبيه سازي فعاليت داده کاوي » ارائه نمود. [4]

عمل داده کاوی از یک پایگاه داده به چند مرحله مشخص تقسیم می شود که ما در این مقاله به معرفی و توضیحی مختصر در مورد هر یک از این مراحل اکتفا می کنیم :

۱. مرحله اول : تشکیل انبار داده .

با توجه به عنوان ، این مرحله برای تشکیل محیطی پیوسته و یک پارچه جهت انجام مراحل بعدی و داده کاوی در آن، انجام می گیرد.در حالت کلی انبار داده مجموعه پیوسته و طبقه بندی شده است که دائما در حال تغییر بوده و دینامیک است که برای کاوش آماده می شود.

۲. مر حله دوم : انتخاب داده ها

در این مرحله برای کم کردن هزینه های عملیات داده کاوی، داده هایی از پایگاه داده انتخاب می شوند که مورد مطالعه هستند و هدف داده کاوی دادن نتایجی در مورد آنهاست.

۳. مرحله سوم : تبدیل داده ها .

مشخص است برای انجام عملیات داده کاوی لزوما باید تبدیلات خاصی روی داده ها انجام گیرد ممکن است این تبدیلات خیلی راحت و مختصر مثل تبدیل byte به integer باشد یا خیلی پیچیده و زمان بر و با هزینه های بالا مثل تعریف صفات جدید و یا تبدیل و استخراج داده ها از مقادیر رشته ای و ... باشد.

٤. مرحله چهارم : کاوش در داده ها .

در این مرحله است که داده کاوی انجام می شود.در این مرحله با استفاده از تکنیک های داده کاوی داده ها مورد کاوش قرار گرفته ، دانش نهفته در آنها استخراج شده و الگو سازی صورت می گیرد.

۵. مر حله پنجم : تفسیر نتیجه .

در این مرحله نتایج و الگو های ارائه شده توسط ابزار داده کاو مورد بررسی قرار گرفته و نتایج مفید معیین می شود.

طرز کار ابزار داده کاو اینگونه است که ابزار به دنبال اثبات این است که وجود چیزی به معنای وجود چیز دیگری است و سعی می کند در درجه اول از توالی ارتباطات برای کشف یک الگو بهره بگیرد و در نهایت اطلاعات بدست آمده را دسته بندی کند تا به الگوی خاصی برسد که بتواند آن را براساس فاکتورهای داخی به مخاطبش ارائه دهد.

همچنین در داده کاوی از الگوریتم های ژنتیک و شبکه های عصبی هم استفاده می شود.شبکه های عصبی به علت کار آمدی در حل مسائل پیچیده و بزرگ مورد استفاده اند و کاربرد الگوریتم های ژنتیک

در داده کاوی برای جستجو و ساختن یک مدل بهینه در میان مدل های بدست آمده است ، به این گونه که مدل های اولیه روی کرومزوم هایی قرار می گیرند و با رقابت بر سر انتقال صفات به نسل بعد ، بهترین مدل و لایق ترین آنها به کاربر ارائه می شوند.

داده کاوی امروز گسترش زیادی یافته است به طوری که اکثر نرم افرار های پایگاه داده ای مثل SQL Server و ORACLE نیز شامل ابزارهایی داده کاوی شده اند ولی هنوز نرم افزار های تخصصی داده کاوی همچون Intelligent Miner , Darwin , Mine Set, Knowledge Studio, Data Mind از مهمترین ابزار های داده کاوی اند.

منابع

1-CHRIS RYGIELSKI, "DATA MINING TECHNIQUES FOR CUSTOMER RELATIONSHIP MANAGEMENT", TECHNOLOGY IN SOCIETY, 2002 .

2- HILL L., "CRM: EASIER SAID THAN DONE", INTELLIGENT ENTERPRISE, 1999

4- Microsoft Visual Studio .Net Documentation

5- Client/Server Survival Guide by Robert Orfali, Dan Harkey, Jeri Edwards

٦- شاه‌سمندي، پرستو «داده‌كاوي در مديريت ارتباط با مشتري» (١٣٨٤)، مجله تدبير شماره ١٥٦.

7- Hand. D.J (1998): "Review of Data mining", The American statistician, 52, 112-118.

8- Jeffery W. Seifert , Analyst in information science and Technology Policy, ' Data Mining : An Overview ' December 2004.

داده کاوی (Data Mining) چیست؟

داده کاوی پل ارتباطی میان علم آمار ، علم کامپیوتر ، هوش مصنوعی ، الگوشناسی ، فراگیری ماشین و بازنمایی بصری داده می باشد. داده کاوی فرآیندی پیچیده جهت شناسایی الگوها و مدل های صحیح، جدید و به صورت بالقوه مفید، در حجم وسیعی از داده می باشد، به طریقی که این الگو ها و مدلها برای انسانها قابل درک باشند.
داده کاوی به صورت یک محصول قابل خریداری نمی باشد، بلکه یک رشته علمی و فرآیندی است که بایستی به صورت یک پروژه پیاده سازی شود .
داده ها اغلب حجیم می باشند و به تنهایی قابل استفاده نیستند، بلکه دانش نهفته در داده ها قابل استفاده می باشد. بنابراین بهره گیری از قدرت فرآیند داده کاوی جهت شناسایی الگوها و مدلها و نیز ارتباط عناصر مختلف در پایگاه داده جهت کشف دانش نهفته در داده ها و نهایتا تبدیل داده به اطلاعات، روز به روز ضروری تر می شود.

مثال تفهیمی در مورد داده کاوی

یکی از نمونه های بارز داده کاوی را می توان در فروشگاه های زنجیره ای مشاهده نمود، که در آن سعی می شود ارتباط محصولات مختلف هنگام خرید مشتریان مشخص گردد. فروشگاه های زنجیره ای مشتاقند بدانند که چه محصولاتی با یکدیگر به فروش می روند .
برای مثال طی یک عملیات داده کاوی گسترده در یک فروشگاه زنجیره ای در آمریکای شمالی که بر روی حجم عظیمی از داده های فروش صورت گرفت، مشخص گردید که مردانی که برای خرید قنداق بچه به فروشگاه می روند معمولا آب جو نیز خریداری می کنند. همچنین مشخص گردید مشتریانی که تلویزیون خریداری می کنند، غالبا گلدان کریستالی نیز می خرند. نمونه مشابه عملیات داده کاوی را می توان در یک شرکت بزرگ تولید و عرضه پوشاک در اروپا مشاهده نمود، به شکلی که نتایج داده کاوی مشخص می کرد که افرادی که کراوات های ابریشمی خریداری می کنند، در همان روز یا روزهای آینده گیره کراوات مشکی رنگ نیز خریداری می کنند.
به روشنی این مطلب قابل درک است که این نوع استفاده از داده کاوی می تواند فروشگاه ها را در برگزاری هوشمندانه فستیوال های فروش و نحوه ارائه اجناس به مشتریان یاری رساند.
نمونه دیگر استفاده از داده کاوی در زمینه فروش را می توان در یک شرکت بزرگ دوبلاژ و تکثیر و عرضه فیلم های سینمایی در آمریکای شمالی مشاهده نمود که در آن عملیات داده کاوی، روابط مشتریان و هنرپیشه های سینمایی و نیز گروه های مختلف مشتریان بر اساس سبک فیلم ها (ترسناک، رمانتیک، حادثه ای و ...) مشخص گردید.
بنابراین آن شرکت به صورت کاملا هوشمندانه می توانست مشتریان بالقوه فیلم های سینمایی را بر اساس علاقه مشتریان به هنرپیشه های مختلف و سبک های سینمایی شناسایی کند.

از دیگر زمینه های به کارگیری داده کاوی، استفاده بیمارستانها و کارخانه های داروسازی جهت کشف الگوها و مدلهای ناشناخته تاثیر دارو ها بر بیماری های مختلف و نیز بیماران گروه های سنی مختلف را می توان نام برد.
استفاده از داده کاوی در زمینه های مالی و بانکداری به شناخت مشتریان پر خطر و سودجو بر اساس معیار هایی از جمله سن ، درآمد، وضعیت سکونت، تحصیلات، شغل و غیره می انجامد.

تعاریف داده کاوی

داده کاوی استخراج اطلاعات مفهومی، ناشناخته و به صورت بالقوه مفید از پایگاه داده می باشد .
Source: W.Frawley and G. Piatetsky. Knowledge Discovery I DataBases.ISSN 0738-4602

داده کاوی علم استخراج اطلاعات مفید از پایگاه های داده یا مجموعه داده ای می باشد .
Source: D. Hand,H. Mannila,P. Smyth(2001).Principles of Data Mining.MIT Press,Cambridge

داده کاوی استخراج نیمه اتوماتیک الگوها، تغییرات، وابستگی ها، نابهنجاری ها و دیگر ساختارهای معنی دار آماری از پایگاه های بزرگ داده می باشد .

Source: R.Grossman

خدمات شرکت دانا پرداز در این زمینه

تفاوت داده کاوی و آنالیز های آماری

داده کاوی معمولا با نوشتن مقدار زیادی گزارش و تحقیق و استعلام در آنها اشتباه گرفته می شود. اما در واقع داده کاوی هیچ کدام از اینها را شامل نمی شود.
داده کاوی توسط تجهیزات خاصی صورت می پذیرد، که عملیات کاوش را بر اساس تجزیه و تحلیل مکرر داده ها انجام می دهد .

داده کاوي با آنالیز های متداول آماری نیز متفاوت است؛در زیرمی توان برخی از اصلی ترین تفاوت های داده کاوي و آنالیز آماری را مشاهده نمود :

آنالیز آماری:

- آمار شناسان همیشه با یک فرضیه شروع به کار می کنند.
- آنها از داده های عددی استفاده می کنند.
- آمارشناسان باید رابطه هایی را ایجاد کنند که به فرضیه آنها مربوط است.
- آنها می توانند داده های نابجا و نادرست را در طول آنالیز مشخص کنند.
- آنها می توانند نتایج کار خود را تفسیر و برای مدیران بیان کنند.

داده کاوی:

- به فرضیه احتیاجی ندارد.
- ابزارهای داده کاوی از انواع مختلف داده ، نه تنها عددی می توانند استفاده کنند.
- الگوریتمهای داده کاوی به طور اتوماتیک روابط را ایجاد می کنند.

- داده کاوی به داده های صحیح و درست نیاز دارد.

- نتایج داده کاوی نسبتا پیچیده می باشد و نیاز به متخصصانی جهت بیان آنها به مدیران دارد.

جهت درک بهتر تفاوت داده کاوی و آنالیزهای آماری به مثال زیر که در مورد شناخت کلاهبرداری های شرکت بیمه می باشد، توجه کنید.

روش آنالیز آماری:

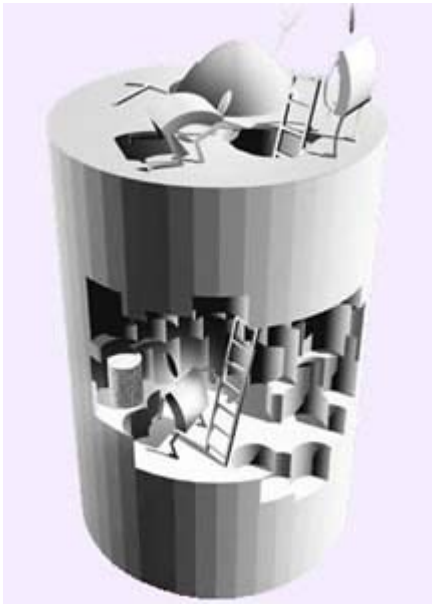
یک مفسر ممکن است متوجه الگوی رفتاری شود که سبب کلاهبرداری بیمه گردد. بر اساس این فرضیه، مفسر به طرح یک سری سوال می پردازد تا این موضوع را بررسی کند. اگر نتایج حاصله مناسب نبود، مفسر فرضیه را اصلاح می کند و یا با انتخاب فرضیه دیگری مجددا شروع می کند. این روش نه تنها وقت گیر است بلکه به قدرت تجزیه و تحلیل مفسر نیز بستگی دارد.
مهمتر از همه اینکه این روش هیچ وقت الگوهای کلاهبرداری دیگری را که مفسر به آنها مظنون نشده و در فرضیه جا نداده ، پیدا نمی کند.

روش داده کاوی:

یک مفسر سیستم های داده کاوی را ساخته و پس از طی مراحلی از جمله جمع آوری داده ها، یکپارچه سازی و اخلاص داده ها به انجام عملیات داده کاوی می پردازد.
داده کاوی تمام الگوهای غیرعادی را که از حالت عادی و نرمال انحراف دارند و ممکن است منجر به کلاهبرداری شوند را پیدا می کند.
نتایج داده کاوی حالت های مختلفی را که مفسر باید در مراحل بعدی تحقیق کند، نشان می دهند. در نهایت مدل های به دست آمده می توانند مشتریانی را که امکان کلاهبرداری دارند، پیش بینی نمایند.

خدمات شرکت دانا پرداز در این زمینه

فواید و نقش داده کاوی در فعالیت شرکتها



امروزه عملیات داده کاوی به صورت گسترده توسط تمامی شرکت هایی که مشتریان در کانون توجه آنها قرار دارند، استفاده می شود، از جمله فروشگاه ها، شرکت های مالی، ارتباطاتی، بازاریابی و غیره .
استفاده از داده کاوی به این شرکتها کمک می کند تا ارتباط عوامل داخلی از جمله قیمت، محل قرارگیری محصولات، مهارت کارمندان را با عوامل خارجی از جمله وضعیت اقتصادی، رقابت در بازار و محل جغرافیایی مشتریان کشف نمایند .
از آنجائیکه هوش مصنوعی یکی از اصلی ترین عناصر داده کاوی
می باشد و با توجه به اینکه به کمک سیستم های کامپیوتری و پایگاه های داده، روزانه به میزان داده ها افزوده می شود، بنابراین استفاده هوشمندانه از دانش بالقوه ای که در این داده نهفته است در دنیای رقابتی امروز برای شرکت ها حیاتی می باشد .

داده کاوی پیش بینی وضع آینده بازار، گرایش مشتریان و شناخت سلیقه های عمومی آنها را برای شرکت ها ممکن می سازد .

مراحل اصلی داده کاوی

داده کاوی را " کشف دانش در داده ها " نیز می نامند. کشف دانش داده ها دارای مراحل مختلفی می باشد که در اینجا به صورت خلاصه آنها را بیان می کنیم :

- استخراج اطلاعات از چندین منبع داده (پایگاه داده).
- یکپارچه سازی اطلاعات و حذف داده های زاید.
- قرار دادن اطلاعات اصلاح شده در انبار داده ها.
- انجام عملیات داده کاوی توسط نرم افزار های مخصوص.
- نمایش نتایج به صورت قابل فهم مانند گزارش و گراف.

انبار داده (Data Warehousing) چیست؟

انبار داده به مجموعه ای از داده ها گفته می شود که از منابع مختلف اطلاعاتی سازمان جمع آوری ، دسته بندی و ذخیره می شود. در واقع یک انبار داده مخزن اصلی کلیه داده های حال و گذشته یک سازمان 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 (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 می باشند که موجب تسریع انجام عملیات آماری و گزارش گیری می شود (OLAP) .

خدمات شرکت دانا پرداز در این زمینه

تاریخچه و دلایل استفاده از انبار داده

از اواخر سال ۱۹۸۰ میلادی، انبار های داده به عنوان نوع متمایزی از پایگاه های داده مورد استفاده اغلب سازمانها و شرکت های متوسط و بزرگ واقع شدند. انبار های داده جهت رفع نیاز رو به رشد مدیریت داده ها و اطلاعات سازمانی که توسط پایگاه های داده سیستم های عملیاتی غیر ممکن بود، ساخته شدند.

سیستمهای عملیاتی سازمان (OLTP) دارای نقاط ضعفی می باشند که انبار های داده آنها را رفع می کنند. از جمله:

- **بار پردازش گزارشات موجب کندی عملکرد برنامه های کاربردی می گردد.**
- **پایگاه های داده برنامه های کاربردی دارای طراحی مناسبی جهت انجام عملیات آماری و گزارش نیستند.**
- **بسیاری از سازمانها دارای بیش از یک برنامه کاربردی (منابع اطلاعاتی) می باشند، بنابراین تهیه گزارشات در سطح سازمان غیر ممکن می شود.**

- **تهیه گزارشات در سیستمهای عملیاتی غالبا نیازمند نوشتن برنامه های مخصوص می باشد که معمولا کند و پرهزینه هستند.**

خدمات شرکت دانا پرداز در این زمینه

مراحل و نحوه ایجاد انبار داده در سازمان

بسیاری از شرکت ها و سازمانها به این باور رسیده اند که گردآوری، سازمان دهی و یکپارچه سازی داده ها در یک مخزن داده برای مدیریت بهینه و اتخاذ تصمیمات کلان یک ضرورت می باشد .

به طور کلی ساخت یک انبار داده، به شکل یک پروژه شامل مراحل اصلی زیر می باشد:

۱- استخراج داده های تراکنشی از پایگاه های داده به یک مخزن واحد

- شناخت منابع داده های سازمان و استخراج داده های ارزشمند از آنها یکی از اصلی ترین مراحل ایجاد انبار داده می باشد .

۲- تبدیل داده ها

- از آنجائیکه سیستمهای اطلاعاتی و برنامه های کاربردی یک سازمان غالبا توسط افراد و پروژه های مختلف به مرور زمان در مواجهه با نیازهای جدید ساخته یا تغییر شکل داده می شوند، یکسان سازی آنها امری ضروری
می باشد.در بسیاری از موارد نیز سیستمهای اطلاعاتی در بستر های مختلف پایگاه داده مانند Microsoft SQL Server ، Oracle، Sybase، Microsoft Accessو غیره طراحی گردیده اند. بررسی جداول، برقراری ارتباط بین فیلدها و یک شکل سازی داده ها در این مرحله صورت می پذیرد .

۳- بارگذاری داده های تبدیل شده به یک پایگاه داده چند بعدی

- بر خلاف پایگاه داده سیستمهای عملیاتی (OLTP) که دارای معماری رابطه ای می باشند و از اصول نرمالیزه استفاده می کنند، طراحی انبار داده به شکلی ویژه بدون بهره گیری از اصول نرمالیزاسیون می باشد. درانبار داده فیلدها در جاهای مختلفی تکرار می شوند و روابط بین جداول کمتر به چشم می خورند. علت آن هم افزایش سرعت پردازش اطلاعات هنگام گزارشات و عملیات آماری می باشد .

٤- تولید مقادیر از پیش محاسبه شده جهت افزایش سرعت گزارش گیری

- مقادیر از پیش محاسبه شده را تراکم نیز می نامند. این مرحله توسط سیستمهایی نظیر Microsoft SQL Server Analysis Services بسیار ساده تر شده است. این تراکم ها که در ابعاد مختلف انبار داده ساخته
می شوند، موجب می شوند که سرعت انجام عملیات گزارش گیری به شکل محسوسی افزایش یابد. بایدتوجه داشت که عملیات ساخت این مقادیر بسیار زمان گیر بوده و نیازمند حافظه زیادی بر روی سروراست .

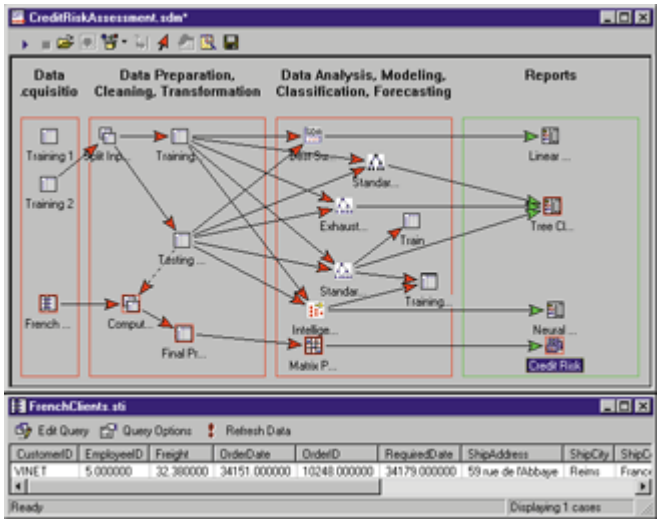
٥- ساخت (یا خرید) یک ابزار گزارش گیری

- پس از انجام مراحل فوق، شما می توانید نسبت به ساخت یا خرید یک نرم افزار گزارش گیری تصمیم گیری نمایید. به طور معمول هزینه ساخت یک نرم افزار گزارش گیری، بالاتر از هزینه خرید آن از یک شرکت خارجی
می شود .

کاوی داده

داده کاوی فرایندی تحلیلی است که برای کاوش داده ها (معمولا حجم عظیمی از داده ها - در زمینه های کسب وکار و بازار) صورت می‌گیرد و یافته‌هابابه‌کارگیری الگوهایی،احراز اعتبار می‌شوند . هدف اصلی داده کاوی پیش بینی است. فرایند داده کاوی شامل سه مرحله می باشد : ۱. کاوش اولیه ۲. ساخت مدل یا شناسایی الگو با کمك احراز اعتبار/ تایید و ۳. بهره برداری.

مرحله ۱ : کاوش. معمولااین‌مرحله با آماده سازی داده ها صورت می گیرد که ممکن است شامل پاك سازی داده ها ،تبدیل داده هاوانتخاب زیرمجموعه‌هايي از رکوردهاباحجم‌عظیمی‌ازمتغییرها(فیلدها) باشد . سپس با توجه‌به‌ماهیت‌مساله تحلیلی، این‌مرحله‌به‌مدل‌هاي پیش بیني ساده یا مدل‌های‌آماری‌وگرافیکی برای شناسایی متغیرهاي مورد نظر و تعیین پیچیدگی مدل‌ها برای استفاده در مرحله بعدی نیاز دارد .



مرحله ۲:ساخت و احراز اعتبار مدل. این‌مرحله‌به بررسی‌مدل‌هاي مختلف و گزینش بهترین مدل با توجه به کارآیی پیش‌بیني آن می پردازد. شاید این مرحله ساده به نظر برسد، اما اینطورنیست. تکنیك‌های‌متعددی‌برای‌رسیدن‌به‌این‌هدف توسعه یافتند.و " ارزیابی رقابتی مدل ها"نام گرفتند. بدین منظور مدل‌های مختلف برای مجموعه داده‌های یکسان‌به‌کارمی‌روندتاکارآیی‌شان‌باهم مقایسه‌شود ،سپس مدلی که‌بهترین کارآیی راداشته باشد، انتخاب می‌شود.این‌تکنیك‌ها عبارتند از : Stacking, Bagging,Boosting و Meta-learning.

مرحله ۳ : بهره برداری. آخرین‌مرحله‌مدلی‌راکه‌درمرحله قبل‌انتخاب‌شده است، در داده‌های‌جدیدبه کارمی‌گیردتا پیش‌بینی‌هاي‌خروجی‌های مورد انتظاررا تولید نماید.داده کاوی‌به‌عنوان‌ابزارمدیریت‌اطلاعات‌برای‌تصمیم گیری،عمومیت‌یافته‌است. اخیرا،توسعه تکنیك های تحلیلی جدید در این زمینه مورد توجه قرار گرفته است (مثلا *Classification Trees*)،اما هنوز داده کاوی مبتنی بر اصول آماری نظیر(*Exploratory Data Analysis (EDA*)می باشد.

بااین وجود تفاوت عمده ای بین داده کاوی و EDA وجوددارد.داده‌کاوی‌بیشتربه‌برنامه‌های‌کاربردی گرایش دارد تا ماهیت اصلی پدیده .به عبارتی‌داده کاوی کمتر با شناسایی روابط بین متغیرها سروکار دارد .

مفاهیم اساسی در داده کاوی

Bagging: این مفهوم برای ترکیب رده بندی های پیش بینی شده از چند مدل به کار می رود.فرض کنیدکه قصداریدمدلی برای رده بندی پیش بیني بسازیدو مجموعه داده های مورد نظرتان کوچك است.شمامی توانید نمونه هایی(با جایگزینی) را از مجموعه داده ها انتخاب و برای نمونه های حاصل ازدرخت رده بندی (مثلا RT&C وCHAID)استفاده نمایید.به طورکلی برای نمونه های مختلف به درخت های متفاوتی خواهید رسید.سپس برای پیش بینی با کمك درخت های متفاوت به دست آمده از نمونه ها ،یك رای گیری ساده انجام دهید.رده بندی نهایی ، رده بندی ای‌خواهد بود که درخت های مختلف آنرا پیش بینی کرده اند .

Boosting: این مفهوم برای تولید مدل‌های چندگانه (برای پیش بینی یا رده بندی)به کار می‌رود. Boosting نیزاز روش RT&C یا CHAID استفاده وترتیبی از classifier ها را تولید خواهد کرد .

Meta-Learning : این مفهوم برای ترکیب پیش بینی‌های حاصل از چند مدل به کار می‌رود.و هنگامی که انواع مدل‌های موجود در پروژه خیلی متفاوت هستند، کاربرد دارد. فرض کنید که پروژه داده کاوی شما شامل Tree classifierها نظیر RT&Cو CHAID، تحلیل خطی و شبکه های عصبی است.هر یك از کامپیوترها،رده بندی هایی رابرای نمونه هاپیش بینی کرده اند.تجربه نشان می‌دهدکه ترکیب پیش بینی های چند روش دقیق تراز پیش بینی های هریك از روشهاست.پیش بینی های حاصل از چند classifier را می توان به عنوان ورودی meta-linear مورد استفاده قرار داد. meta-linear پیش بینی هارا ترکیب می کند تا بهترین رده بندی پیش بینی شده حاصل شود.

يكپارچه سازی، داده كاوی و طراحی مخازن داده

توسعه بكارگيري سيستمهاي اطلاعاتي در سازمانها و گسترش انتقال فرآيندهاي سازماني به سامانه‌هاي الكترونيكي و نيز وابستگي روزافزون مديران و تصميم سازيها به اطلاعات و سيستمهاي اطلاعاتي، مباحث نويني نظير يكپارچه سازي (Integration)، داده كاوي (Data mining) و مخزن داده (Warehouse Data) را مطرح ساخته است.

مديران و تصميم‌گيران سازمانها در وهله اول نيازمند دسترسي به داده ها در هر محل و در مرحله بعد به تحليل آنها مي‌باشند تا بتوانند به مزيت رقابتي در بازار در مقابل رقيبان دست يابند. اين تحليل‌ها شامل تشخيص خوشه‌بندي در داده و يافتن تمايلات مشتريان است كه از حوزه پايگاه داده‌هاي معمولي خارج مي‌باشد. براي انجام اين مهم نه تنها داده بلكه سابقه آن نيز ضروري است. مخزن داده براي اينگونه تحليل‌ها قادر است داده را از منابع مختلف روي سيستم عامل‌هاي متفاوت و غيره جمع‌آوري و خلاصه‌بندي كند و بعد از فراهم سازي مخزن داده ابزارهايي براي كاربر نهايي بايد فراهم گردد تا بتواند بنحو مطلوب و دلخواه از آن استفاده كند. پردازش تحليلي برخط (OLAP) و داده كاوي اين ابزارها را فراهم مي‌كند. در اين كارگروه عمليات تحليل و طراحي مخزن داده و نهايتاً پياده‌سازي آن انجام گرفته، ابزارهاي مختلف داده كاوي و OLAP براي تحليل تدوين مي‌شود. اين كارگروه تاكنون تجربه كاري متعدد درباره OLAP با استفاده از امكانات سامانه Oracle j و همچنين تجربه پژوهشي در زمينه داده كاوي عمدتاً بر پايه مدلسازي پيش‌بيني مبتني بر سيستمهاي يادگيري را دارا مي‌باشد.

هم اكنون در هر كشور، سازمان يا شركت براي امور بازرگاني، پرسنلي، آموزشي، آماري و . . . پايگاه داده ها ايجاد يا خريداري شده است، به طوري كه اين پايگاه داده ها براي مديران، برنامه ريزان و پژوهشگران. جهت تصميم گيري هاي راهبردي، تهيه گزارش هاي مختلف، توصيف وضعيت جاري مي تواند مفيد باشد. داده كاوي يا استخراج و كشف سريع و دقيق اطلاعات با ارزش و پنهان از اين پايگاه داده ها از جمله اموري است كه هر كشور، سازمان و شركتي به منظور توسعه علمي، فني و اقتصادي خود به آن نياز دارد.
در كشور ما نيز سازمان ها، شركت ها و مؤسسات دولتي و خصوصي از جمله سازمان تامين اجتماعي به طور فزاينده اقدام به ايجاد يا خريد نرم افزارهاي پايگاه داده ها و مكانيزه كردن سيستم هاي اطلاعات خود هستند، همچنين با توجه به فصول دهم و يازدهم قانون برنامه سوم توسعه در خصوص داد و ستدهاي الكترونيكي و همچنين تأكيد بر برخورداري كشور از فن آوري هاي جديد اطلاعات براي دستيابي آسان به اطلاعات داخلي و خارجي، دولت مكلف شده است امكانات لازم براي دستيابي آسان به اطلاعات، زمينه سازي براي اتصال كشور به شبكه هاي جهاني و ايجاد زير ساخت هاي ارتباطي و شاهراه هاي اطلاعاتي فراهم كند. واضح است اين امر باعث ايجاد پايگاه هاي عظيم داده ها شده و ضرورت استفاده از داده كاوي را كه فرآيند خودكار كشف دانش و اطلاعات از پايگاه هاي داده مي باشد، بيش از پيش نمايان مي سازد.
هدف از داده كاوي ايجاد مدل هايي براي تصميم گيري است. اين مدلها رفتارهاي آينده را براساس تحليلهاي گذشته پيش بيني مي كنند. به كاربردن داده كاوي به عنوان اهرمي براي آماده سازي داده ها و تكميل قابليتهاي انباره داده، بهترين موقعيت را براي به دست آوردن برتريهاي رقابتي يا خدمت رساني بهتر به مشتري ايجاد مي كند.
با توجه به اينكه سازمان تامين اجتماعي و ديگر سازمانهاي بيمه ديگر مانند سازمان بيمه خدمات درماني، سازمان بازمشستگي كشوري و ... همواره با داده‌ها و اطلاعات بسيار زيادي در مورد سوابق بيمه شده ها، كارفرمايان، اطلاعات درماني بيمه شده ها، اطلاعات پرسنل، منابع مادي و... روبروهستند و در اكثر مواقع اين داده‌ها مي‌تواند حامل اطلاعات و الگوهاي باارزشي باشند، لذا يكي از مهمترين كاربردهاي داده‌كاوي در اينگونه سازمانها است.
امروز بانك‌هاي اطلاعاتي وسيعي از ويژگي‌هاي بيمه شده ها در سازمان تامين اجتماعي موجود است كه اطلاعات مربوط به ويژگي‌هاي خانوادگي، تحصيلي و ... را شامل مي‌شود. يافتن الگوها و دانش نهفته در اين اطلاعات به تصميم‌گيرندگان در اين زمينه كمك شاياني خواهد كرد. استفاده از تكنيك‌هاي

پيشرفته داده‌كاوي مانند خوشه‌بندي، طبقه‌بندي، و ... مي‌تواند در طبقه‌بندي كارفر مايان و كارگاه ها، مراكز درماني، يافتن الگوهاي خاص و با ارزش در مورد مراكز درماني موفق، يافتن استراتژي توزيع تجهيزات درماني، منابع مالي يافتن نقاط بحراني در مديريت مالي و موارد ديگر كاربرد داشته باشد.
داده كاوي مي تواند براي پاسخ دادن به يک سوال خاص مربوط به بيمه شده و نيز براي كشف روندهاي عمومي كه به تصميم گيري كمک مي كنند، استفاده شود . براي مثال سوال مي تواند چنين باشد: امكان اينكه بيمه شده بعد از ده سال از شروع بيمه از كار افتاده شود چقدر است؟ يا ميزان اعتبار مورد انتظار براي مستمري بگيران سازمان يا اعتبار لازم براي درمان در سال آينده چقدر است؟ درک الگوي استفاده كلي از خدمات درماني يا تحليل درخواستهاي براي طي ٥ سال گذشته نيز همگي مثالهايي از كشف روندهاي عمومي اند.
هم اينک در شركت مشاور مديريت و خدمات ماشيني تامين دانش و شناخت كافي بر روي شيوه ها و ابزار هاي داده كاوي فرآهم شده است و در اين زمينه و يافتن الگوهاي متفاوت بر مبناي اطلاعات جمع آوري شده فعاليت مي شود.

Mining Introduction of Data¿

چهار شنبه ۱۶ آذر ۱۳۸۴

کاربردهای داده کاوی در کتابخانه ها و موسسات دانشگاهی*

نوشته ویوک.ان. پاتکار(V. N. Patkar)
ترجمه مریم صراف زاده و افسانه حاضری
Email: mmsarraf@yahoo.com
دانشجویان دکتری سیستمهای اطلاعاتی- ملبورن استرالیا

چکیده:
کتابخانه ها و موسسات آموزشی با مشکل مدیریت کارآمد بار سنگین داده ها که دائما نیز در حال افزایش است روبرو می باشند. نرم افزارهای کامپیوتری بکار گرفته شده برای این منظور، غالبا فقط برای پرس و جوهای معمولی و پشتیبانی از مسائل مدیریتی و برنامه ریزی کوتاه مدت اداری جوابگو هستند. در حالیکه در عمق درون این حجم داده ها، الگوها و روابط بسیار جالبی میان پارامترهای مختلف بصورت پنهان باقی میماند. داده کاوی یکی از پیشرفتهای اخیر در حوزه کامپیوتر برای اکتشاف عمیق داده هاست. داده کاوی از اطلاعات پنهانی که برای برنامه ریزیهای استراتژیک و طولانی مدت میتواند حیاتی باشد پرده برداری میکند. تبیین مشخصه های اساسی فرایند داده کاوی و کشف کاربردهای ممکن آن در کتابداری و موسسات دانشگاهی اهداف اصلی این مقاله را شکل میدهند.

مقدمه
در دنیای بشدت رقابتی امروز، اطلاعات بعنوان یکی از فاکتورهای تولیدی مهم پدیدار شده است. در نتیجه تلاش برای استخراج اطلاعات از داده ها توجه بسیاری از افراد دخیل در صنعت اطلاعات و حوزه های وابسته را به خود جلب نموده است.
حجم بالای داده های دائما در حال رشد در همه حوزه ها و نیز تنوع آنها به شکل داده متنی، اعداد، گرافیکها، نقشه ها، عکسها، تصاویر ماهواره ای و عکسهای گرفته شده با اشعه ایکس نمایانگر پیچیدگی کار تبدیل داده ها به اطلاعات است. علاوه بر این، تفاوت وسیع در فرآیندهای تولید داده مثل روش آنالوگ مبتنی بر کاغذ و روش دیجیتالی مبتنی بر کامپیوتر، مزید بر علت شده است.
استراتژیها و فنون متعددی برای گردآوری، ذخیره، سازماندهی و مدیریت کارآمد داده های موجود و رسیدن به نتایج معنی دار بکار گرفته شده اند. بعلاوه، عملکرد مناسب ابرداده[۱] که داده ای درباره داده است در عمل عالی بنظر میرسد.

پیشرفتهای حاصله در علم اطلاع رسانی و تکنولوژی اطلاعات، فنون و ابزارهای جدیدی برای غلبه بر رشد مستمر و تنوع بانکهای اطلاعاتی تامین می کنند. این پیشرفتها هم در بعد سخت افزاری و هم نرم افزاری حاصل شده اند. ریزپردازنده های سریع، ابزارهای ذخیره داده های انبوه پیوسته و غیر پیوسته، اسکنرها، چاپگرها و دیگر ابزارهای جانبی نمایانگر پیشرفتهای حوزه سخت افزار هستند. پیشرفتهای حاصل در نظامهای مدیریت بانک اطلاعات در طی چهار دهه گذشته نمایانگر تلاشهای بخش نرم افزاری است. این تلاشها در بخش نرم افزار را میتوان بعنوان یک حرکت

پیشرونده از ایجاد یک بانک اطلاعات ساده تا شبکه ها و بانکهای اطلاعاتی رابطه ای و سلسله مراتبی برای پاسخگویی به نیاز روزافزون سازماندهی و بازیابی اطلاعات ملاحظه نمود. بدین منظور در هر دوره، نظامهای مدیریت بانک اطلاعاتی[۲] مناسب سازگار با نرم افزار سیستم عامل و سخت افزار رایج گسترش یافته اند. در این رابطه میتوان از محصولاتی مانند، Dbase-IV, Unify, Sybase, Oracle و غیره نام برد.

داده کاوی یکی از پیشرفتهای اخیر در راستای فن آوریهای مدیریت داده هاست. داده کاوی مجموعه ای از فنون است که به شخص امکان میدهد تا ورای داده پردازی معمولی حرکت کند و به استخراج اطلاعاتی که در انبوه داده ها مخفی و یا پنهان است کمک می کند. انگیزه برای گسترش داده کاوی بطور عمده از دنیای تجارت در دهه ۱۹۹۰ پدید آمد. مثلا داده کاوی در حوزه بازاریابی، بدلیل پیوستگی غیرقابل انتظاری که بین پروفایل یک مشتری و الگوی خرید او ایجاد میکند اهمیتی خاص دارد. (Barry and Linoff, 1997)

تحلیل رکوردهای حجیم نگهداری سخت افزارهای صنعتی، داده های هواشناسی و دیدن کانالهای تلوزیونی از دیگر کاربردهای آن است. در حوزه مدیریت کتابخانه کاربرد داده کاوی بعنوان فرایند ماخذ کاوی[۳] نامگذاری شده است. این مقاله به کاربردهای داده کاوی در مدیریت کتابخانه ها و موسسات آموزشی می پردازد. در ابتدا به چند سیستم سازماندهی داده ها که ارتباط نزدیکی به داده کاوی دارند می پردازد؛ سپس عناصر داده ای توصیف میشوند و درپایان چگونگی بکارگیری داده کاوی در کتابخانه ها و موسسات آموزشی مورد بحث قرار گرفته و مسائل عملی مرتبط در نظر گرفته می شوند.

پیشرفت در تکنولوژیهای داده پردازی

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

۱. رکوردهای اصلی [٤]
۲. رکوردهای عملیاتی[٥]

فرض بر این است که رکوردهای اصلی حاوی اطلاعات پایه هستند که معمولا چندان تغییر نمی کنند در حالیکه رکوردهای عملیاتی با توجه به طبیعت عملیات تجاری حتی بطور ساعتی تغییر خواهند کرد.

سیستمهای مدیریت پایگاه داده[٦] مناسب برای پیوند دادن این دو مجموعه اطلاعاتی و تهیه گزارشهای استاندارد جهت کنترل فعالیتها گسترش یافتند. سیستم اطلاعات مدیریت رایج برای پشتیبانی عملیات و سرویس دهی به چند کاربر در سطوح مختلف سازمان مبتنی بر این نظریه است.

بمنظور کمک به تصمیم گیری راهبردی، نظریه تاسیس بانک اطلاعات رکوردهای اصلی به نظریه سازماندهی دیتا مارت[۷] و انبار داده ها[۸] تغییر یافت. استخراج اطلاعات از رکوردهای عملیاتی یا پایگاههای اطلاعات عملیاتی و سازماندهی آن برای تحلیل استاندارد یا زمانی فلسفه اولیه و اصولی چنین پیشرفتهایی است. گرچه، دیتا مارت و انبار داده ها از نظر هدف و ساختار با هم

متفاوتند. (Inmon, ۱۹۹۸)

دیتامارت

دیتا مارت اغلب کوچک است و بر یک موضوع یا دپارتمان خاص متمرکز است. بنابراین پاسخگوی یک نیاز داخلی است. طرح بانک اطلاعات برای یک دیتامارت حول ساختار اتصال ستاره ای ساخته شده است که بهینه برای نیازهای کاربران دپارتمان است. دیتامارت معمولا با ابزارهای کامپیوتری که انعطاف پذیری تحلیل را تامین میکنند اما ممکن است برای سازماندهی حجم بالای داده ها مناسب نباشند؛ نیرومند میشود. رکوردهای ذخیره شده در دیتامارتها بخوبی نمایه شده اند.

یک دیتامارت در صورتیکه داده ها را از منابع داده ای بسیار سازماندهی شده مثل انبار داده ها بگیرد؛ دیتامارت وابسته نامیده میشود. مسلما دیتامارتهای وابسته از لحاظ ساختاری و معماری منطقی هستند. منبع دیتامارتهای وابسته تکنولوژی بانک اطلاعات دپارتمانی است. دیتامارتهای مستقل ثابت نیستندو از لحاظ معماری بسیار با هم متفاوتند. این مساله هنگام یکپارچه سازی دیتامارتهای مستقل، مشکل ایجاد میکند. بنابراین با یکپارچه سازی ساده دیتامارتها یک انبار داده ایجاد نخواهد شد.

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

انبار داده ها

یک انبار داده کاملا " متفاوت از دیتامارت است. سازماندهی انبارهای داده بگونه ایست که کلیه موضوعات حول فعالیتهای کاری سازمان را می پوشاند. انبار داده نمایانگر یک تسهیلات مرکزی است.

برخلاف دیتامارت که در آن داده ها به شکل خلاصه تر و متراکم تر وجود دارند، یک انبار داده ، داده ها را در یک سطح نامتراکم ذخیره می کند. ساختار داده ها در یک انبار داده یک ساختار لزوما" هنجار شده است. بدین معنی که ساختار و محتوای داده ها در انبار داده منعکس کننده ویژگیهای دپارتمانهای عضو نیست. داده ها در انبار داده از نظر حجم و شکل کاملا" متفاوت از داده ها در دیتامارت هستند. دیتامارت ممکن است شامل حجم زیادی از داده های قدیمی و گذشته نگر باشد. داده ها در انبار داده اغلب بصورت نسبتا" سبک نمایه میشوند. (به بیان دیگر در عمق کمتر).

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

وابستگی به زمان:

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

غیر فرار بودن[۹]:

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

ثبات داده ها را برای استفاده های وسیعتر تضمین میکند.

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

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

واضح است که انبار داده اساسا" برای پرس و جوهای پشتیبان تصمیم گیری ساخته شده است. بر این اساس سازماندهی وعملیات انبار داده چنان طراحی شده اند تا نیازهای اطلاعاتی روزمره یا معمولی را پاسخگو باشند. بدلیل حجم بسیار بالای چنین پایگاه اطلاعاتی یک سیستم کامپیوتری پیشرفته برای عملیات انبارسازی داده ها لازم است. همچنین یک بانک اطلاعات مجزا شامل ابرداده که مشخصه هایی نظیر نوع، فرمت، مکان و پدیدآورندگان داده های ذخیره شده در یک انبار داده ها را توصیف میکند نیز برای کمک به کاربران و مدیران داده ها ساخته میشود. مشخص شد که انبار داده بدلیل اندازه و تنوعش، اگر مبتکرانه پردازش شود میتواند به تولید اطلاعاتی منجر شود که در وهله اول آشکار نیستند. با انتخاب متناسب داده ها، بکار گرفتن فنون مختلف غربال کردن و تفسیر زمینه ای [۱۰]، داده ذخیره شده میتوانست منجر به کشف الگوها یا رابطه هایی شود که بینش نویی به تصمیم گیرنده دهد. این مساله نظریه توسعه عملیات داده کاوی را به موازات معدن کاوی بروز داد. ذکر این نکته لازم است که داده کاوی در اصل لزوما" نیاز به سازماندهی یک انبار داده ندارد. حال به داده کاوی می پردازیم.

عناصر داده کاوی

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

۱. انتخاب داده ها
۲. پاک سازی داد ها
۳. غنی سازی داده ها
٤. کد گذاری داده ها

با دارا بودن هدف کلی در مطالعه، انتخاب مجموعه داده های اصلی برای تحلیل، اولین ضرورت است. رکوردهای لازم میتواند از انبار داده ها و یا بانک اطلاعاتی عملیاتی استخراج شود. این رکوردهای داده جمع آوری شده؛ اغلب از آنچه آلودگی داده ها نامگذاری شده است رنج می برند و

بنابراین لازم است پاکسازی شوند تا از یکدستی فرمت (شکلی) آنها اطمینان حاصل شود، موارد تکراری حذف شده و کنترل سازگاری دامنه بعمل آید. ممکن است داده های گردآوری شده از جنبه های خاصی ناقص یا ناکافی باشند. در این صورت داده های مشخصی باید گردآوری شوند تا بانک اطلاعات اصلی را تکمیل کنند. منابع مناسب برای این منظور باید شناسایی شوند. این فرایند مرحله غنی سازی داده ها را تکمیل میکند. یک سیستم کدگذاری مناسب معمولا" جهت انتقال داده ها به فرم ساختار-بندی شده جدید؛ متناسب برای عملیات داده کاوی تعبیه میشود .

فنون داده کاوی

ممکن است متوجه شده باشید که فنون داده کاوی یک گروه نامتجانس را شکل میدهند چرا که هر تکنیکی که بتواند بینش جدیدی از داده ها را استخراج کند میتواند داده کاوی به حساب آید. برخی از ابزارهای رایج بکار گرفته شده تحت عنوان داده کاوی عبارتند از: (Adriaans and Zantinge, 2003)

ابزارهای پرس و جو[۱۱]: ابزارهای متداول زبان پرس و جوی ساختاربندی شده[۱۲]در ابتدا برای انجام تحلیلهای اولیه بکار گرفته شدند که می تواند مسیرهایی برای تفحص بیشتر نشان دهد.

فنون آماری: مشخصات اصلی داده ها لازمست با کاربرد انواع مختلفی از تحلیلهای آماری شامل جدول بندی ساده[۱۳] و متقاطع[۱٤] داده ها و محاسبه پارامترهای آماری مهم بدست آید.

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

پردازش تحلیلی پیوسته[۱۵]: از آنجا که مجموعه داده ها ممکن است روابط چندین بعدی داشته باشند، روشهای متعددی برای ترکیب کردن آنها وجود دارد. ابزارهای پردازش تحلیلی پیوسته به ذخیره چنین ترکیباتی کمک میکند و ابزارهای ابتدا-انتها[۱٦]ی پیوسته برای انجام پرس و جو ایجاد میکند. اما این ابزارها هیچ دانش جدیدی ایجاد نمی کنند.

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

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

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

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

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

گام نهایی فرایند داده کاوی، گزارش دادن است. گزارش شامل تحلیل نتایج و کاربردهای پروژه، درصورت بکارگیری آنها، است . و متن مناسب، جداول و گرافیکها را در خود جای می دهد. بیشتر اوقات گزارش دهی یک فرایند تعاملی است که تصمیم گیرنده با داده ها در پایانه کامپیوتری بازی میکند و فرم چاپی برخی نتایج واسطه محتمل را برای عملیات فوری بدست می آورد.
داده کاوی در تولید چهار نوع دانش ذیل مفید است: ((Fayyad et al., 1996
- دانش سطحی (کاربردهای (SQL
- دانش چند وجهی (کاربردهای (OALP
- دانش نهان (تشخیص الگو و کاربردهای الگوریتم یادگیری ماشینی)
- دانش عمیق (کاربردهای الگوریتم بهینه سازی داخلی)

نرم افزار:
از آنجا که داده کاوی با بانکهای اطلاعاتی بزرگ سروکار دارد، به گونه ای ایده ال با تکنولوژی خدمت گیر-خدمت گر[۱۷] بکار میرود. کاربردهای عمومی داده کاوی بیشتر شامل تقسیم کردن داده ها در خوشه های مقتضی، کدگذاریهای مناسب، کاوش برای الگوها و طراحی کردن با استفاده از فنون آماری و الگوریتمهای ژنتیکی است. تعداد زیادی از بسته های نرم افزاری واجد این جنبه های ابزارهای داده کاوی با درجات متفاوتی از جامعیت در دسترس هستند. برای مثال بسته های نرم افزاری که منحصرا" برای کاربردهای OLAP در دسترس هستند عبارتند از: Oracle OLAP, DB2 OLAP Server, CleverPath OLAP . نرم افزارهای آماری عمومی مثل SPSS, SAS, STATISTICA با امکاناتی برای داده کاوی و بسته های نرم افزاری اختصاصی داده کاوی مثل Weka, Insightful Miner3, Text Mining Software, Enterprise Data Mining software, PolyAnalyst 4.6 مفید هستند.

کاربردهای داده کاوی در کتابخانه ها و محیط های دانشگاهی

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

مدیریت و خدمات کتابخانه

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

جدول یک- کاربردهای داده کاوی در کتابخانه ها

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

داده کاوی میتواند برای پاسخ دادن به یک سوال خاص مربوط به کتابخانه و نیز برای کشف روندهای عمومی که به تصمیم گیری کمک میکنند، استفاده شود . برای مثال سوال میتواند چنین باشد: امکان اینکه امانت گیرندگان منابع را یک هفته بعد از تاریخ عودت برگردانند تا نامه های یادآوری کمتری فرستاده شود چقدر است؟ یا میزان اشتراک مورد انتظار برای نشریات بین المللی انتخاب شده برای سال آینده چقدر است؟ درک الگوی استفاده کلی مجلات الکترونیکی یا تحلیل درخواستهای اعضا برای میکروفیلمها طی ۵ سال گذشته نیز همگی مثالهایی از کشف روندهای عمومی اند. دامنه تحلیل استنادی هم میتواند با استفاده از داده کاوی گسترش داده شود.
در ارتباط با کتابخانه ها، وب کاوی حوزه دیگری از علاقمندی است. وب کاوی شامل محتوا کاوی وب، ساختار کاوی وب و استفاده کاوی وب با توجه به یک موضوع خاص است که در طراحی خدمات جدید مبتنی بر وب کمک خواهد کرد.
مدیریت موسسات دانشگاهی
اداره موسسات دانشگاهی کار پیچیده ای است. در این موسسات دائما" نیاز به درآمدزایی و خود-کارآمدی و کاهش وابستگی به بودجه دولتی احساس میشود. این مساله کنترل دائمی جنبه های مختلف هر فعالیت و پروژه را می طلبد. بانکهای اطلاعاتی برای چنین موسساتی مربوط به دانشجویان، دانشکده، اساتید و کارمندان، تعداد رشته ها و چند مورد دیگر است . ارزیابی تقاضا و وضعیت عرضه نقش مهمی بازی میکند. مرور بانکهای اطلاعاتی نمونه در جدول ۲ نمایانگر کاربردهای بالقوه داده کاویست.

جدول ۲- کاربردهای داده کاوی در موسسات دانشگاهی

بانک اطلاعاتی	کاربرد متصور

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

کاربرد داده کاوی در دانشگاه ملی سنگاپور قابل ملاحظه است. در این دانشگاه از ابزارهای داده کاوی برای شناسایی و دسته بندی دانشجویانی که به کلاسهای پیش نیاز برای واحد درسی ارائه شده نیاز داشتند استفاده شد. (Kurian and John, 2005)
علاوه بر آن، مسائلی مانند اختصاص بهتر منابع و نیروی انسانی، مدیریت روابط دانشجو و به تصویر کشیدن رفتار گروههای مختلف میتواند بوسیله ابزارهای داده کاوی انجام شود.

محدودیت ها
کاربرد داده کاوی با چند عامل محدود شده است. اولین مورد به سخت افزار و نرم افزار لازم و موقعیت بانک اطلاعاتی مربوط میشود . برای مثال در هند، داده های غیر مجتمع که برای کاربردهای داده کاوی لازم است ممکن است به فرم دیجیتالی در دسترس نباشد. در دسترس بودن نیروی انسانی ماهر در داده کاوی نیز مسأله مهم دیگری است. محرمانه بودن رکوردهای مراجعان ممکن است در نتیجه پردازش داده های مبتنی بر داده کاوی آسیب پذیر شود. کتابداران و مؤسسات آموزشی باید این مسأله را در نظر داشته باشند؛ چرا که در غیر اینصورت ممکن است گرفتار شکایات قانونی گردند.
محدودیت دیگراز ضعف ذاتی نهفته در ابزارهای نظری ناشی میگردد. ابزارهایی مانند یادگیری ماشینی و الگوریتمهای ژنتیکی بکار گرفته شده در فعالیتهای داده کاوی به مفاهیم وفنون منطق و آمار بستگی دارد. در این حد نتایج به روش مکانیکی تولید شده و بنابراین به یک بررسی دقیق نیاز دارند. اعتبار الگوهای بدست آمده به این طریق؛ باید آزمایش شود. چرا که که در بسیاری موارد روابط علل و معلول مشتق شده؛ از برخی استدلالات غلط ذیل رنج میبرند. (Cannavo, 2003)

- علت دور

مثلا" امکانات ضعیف خوابگاه باعث می شود دانشجویان نمرات پایینی کسب نمایند

- علت مجرد

مثلا" بودجه محدود بر بازدهی پژوهشی دانشکده تأثیر می گذارد

- علائم در نظر گرفته شده برای این عوامل

مثلا" مجموعه کتابخانه ممکن است افزایش نیابد چون تعداد خوانندگان مرتبا" کاهش می یابد.

- سفسطه دسته بندی

مثلا" مدرسان حقوق بسیار بالا دریافت می کنند و کل حقوقشان بالغ بر میلیونها میشود.

- سفسطه ترکیب

مثلا" اگر هر مدرس در دانشکده شایسته و واجد صلاحیت باشد کل دانشکده عملکرد بهتری خواهد داشت.

- سوگیری در انتخاب نمونه:

مثلا"استناد به یافته های یک پیمایش نمونه گیری شده از دانشجویان یک دانشکده که از خانواده های ثروتمند هستند و مخارج روزانه در خوابگاه برای هر دانشجو ۱۰۰ دلار است.

از آنجایی که مطالعه الگوها و استخراج روابط میان رکوردها مستلزم کاربرد منطق قیاسی و استقرایی است فرد باید مراقب اشتباهاتی که عموما" رخ میدهد باشد. برای مثال بحثهای قیاسی یا استقرایی، تا زمانیکه وضعیت درست بودن فرضیه آزمایش نشود چیزی درباره درست یا غلط بودن نتایجشان نمی گویند. طبیعتا، نتایج تولید شده ماشینی ممکن است از چنین نقایصی رنج ببرند.

تذکرات نهایی

بکارگیری تکنولوژی اطلاعات توسط هر سازمان در عمل یک فرایند هموار نیست. کتابخانه یا مؤسسه دانشگاهی از این قضیه استثنا نیست. اما، تجربه نشان میدهد که یک برنامه نظام مند میتواند ظهور و نگهداری تکنولوژی اطلاعات در محیط کتابخانه را تسهیل کند. (Patkar and Iyer, 1990; Patkar, ۲۰۰۰, ۲۰۰٤) حتی کاربرد تکنولوژی های پیشرفته پردازش اطلاعات مثل سیستمهای خبره و سیستم اطلاعات جغرافیایی (جی.آی.اس) در کتابخانه گزارش شده است. (Patkar, 1999 ;Myers, 1992) با این پیش زمینه ، کاربرد داده کاوی بوسیله کتابخانه ها و موسسات دانشگاهی ، به شرط آماده سازی مناسب، بطور قابل توجهی عملی است.

برای دانشگاهها، کالجها، مدارس و موسسات آموزش از راه دور که بانکهای اطلاعاتی عظیمی دارند، ابزارهای داده کاوی میتواند الگوها و روابطی را که خیلی عیان نیستند آشکار کند. این نتایج ممکن است به طراحی دوباره فرایندها و رویه های مرتبط منجر شود. تحلیلهای پشتیبانی شده توسط داده کاوی در کل موسسات و محیط ها میتواند مسائل متنوع مدیریت آموزشی؛ از جمله درک بهتر مشخصه های اقتصادی اجتماعی دانشجویان، مندرجات رشته ها و آموزش و پرورش و ساختار هزینه را مخاطب قرار دهد.

آنچه لازم است اینست که فراتر از عملکرد داده پردازی استاندارد قدم برداریم مخصوصا" کتابخانه ها و موسسات دانشگاهی که با انواع مختلفی از بانکهای اطلاعاتی سروکار دارند و به سطوح معقولی از کامپیوتری کردن و دیجیتالی کردن داده ها دست یافته اند. در یک نظر، ابزارهای داده کاوی نمایانگر پیشرفت در زنجیره تکنولوژی اطلاعات هستند. داده کاوی همچنین میتواند بعنوان بخشی از فرایند بزرگتر کشف دانش در بانکهای اطلاعاتی در محیط های مختلف در نظر گرفته شود. البته نباید چنین پنداشت که ابزارهایی مثل داده کاوی نیاز به مداخله انسانی را کاهش خواهد داد. همچنانکه در بالا نشان داده شد، ارزیابی و تعدیل نتایج بدست آمده بوسیله چنین ابزارهای خودکاری؛ به آزمایش نیاز دارد تا در برابر کاربردهای غلط محافظت شود.

انتظار میرود داده کاوی در گسترش سازمان خودیادگیرنده مشارکت کند. کشف انتخابهای نوین با بهره گیری از داده کاوی اطمینان بخش بهترین کاربرد ممکن منابع موجود است. داده کاوی ماهیت چرخه مانند دارد. برای اینکه در پی کشف الگوها، سوالات بیشتری پدید خواهند آمد که دور بعدی فرایند را شکل میدهند. بهره برداری از تکنولوژیهای پیشرفته مثل داده کاوی مطمئنا" برای

متخصصان کتابداری و مدیران موسسات آموزشی یک چالش دائمی خواهد بود ؛ چرا که آنها خلاقیت طلبند و برای نوآوری تلاش می کنند.

!Error

REFERENCES

1. Adriaans, P; Zantinge, D. *Data Mining. 9th Indian reprint*. Delhi: Pearson Education, 2003.
2. Barry, M; Linoff, G. *Data Mining Techniques for Marketing, Sales and Customer Support*. New York: John Wiley & Sons, 1997.
3. Cannavo, S. *Think to Win: The Power of Logic in Everyday Life*. Mumbai: Magna Publishing Co. Ltd. 2003.
4. Fayyad, U.M; Piatetsky-Shapiro G; Smyth, P; Uthurusamy, R. *Advances in Knowledge Discovery and Data Mining*. Cambridge MA: AAAI Press/MIT Press, 1996
5. Inmon, B. Data Mart does not Equal Data Warehouse. *DM Review*. May 1998. (http://www.dmreview.com/article_sub.cfm?articleId=608).
6. Kurian, J.C; John, B.M. *Mining the Education Domain*. http://www.getforme.com/previous 2004/151004_miningtheeducationdomain.htm, (accessed on March 31, 2005).
7. Myers, J.E. Reference Expert: Developing a Computer Expert System for Library Reference Service. Paper presented at the *Conference on Advances in IT Applied to Libraries*, University of Puerto Rico, Rio Piedras Campus, 8 October 1992.
8. Nicholson, S. The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making. *Information Technology and Libraries*. Vol.22(4); 2003; p4-9.
9. Patkar, V.N; Iyer, P.V. Determining Priorities for Computerisation in Libraries. *Annals of Library Science and Documentation. Vol.* 37(3); 1990; p110-116.
10. —. Application of GIS in Library Management. *Information Studies*. Vol. 5(2); 1999; p73-82.
11. —. Information Technology Adoption by Libraries: Opportunities and Barriers. *Information Studies*. Vol. 6(2); 2000; p125-140.
12. —. Adaptation to Information Technology Development by Libraries. *Information Studies*. Vol. 10(4); 2004; p637-648.

یادداشتها:

[1] Metadata

[2] DBMS

[3] bibliomining

[4]Master records

[5] Transaction records

[6] DBMS=Database Management System

[7] Data Mart

[8] Data warehouse

[9] Non-volatility

[10]Contextual interpretation

[11] Query tools

[12] Structured Query Language(SQL)

[13] Simple tabulation

[14] Cross tabulation

[15] Online analytical processing (OALP)

[16] Front-end tools

[17] Client-server

---

* این مقاله ترجمه ای است از:

Patkar, V. N. (2005). "Data mining applications in library and academic institutions." Information Studies 11(3): 145-156.

داده كاوي درمديريت ارتباط بامشتري

پرستو شاه سمندي
Parastoushahsamandi@yahoo.com

چكيده

شركتهاي امروزي از طريق تجزيه و تحليل چرخه زندگي مشتري به افزايش ارزش مشتري دست يافته اند. ابزارها و فناوريهاي انبار داده، داده كاوي و ديگر تكنيك هاي مديريت ارتباط با مشتري، روشهايي هستند كه فرصتهاي جديدي را براي تجارت فراهم كرده اند.

در واقع ديدگاه محصول محوري جاي خود را به مشتري محوري داده است. بنابراين، با جمع آوري داده هاي مربوط به مشتري و تصميم گيري براساس الگوهاي استخراج شده از روابط پنهان ميان داده ها به وسيله ابزار داده كاوي، مي توان به خواسته مشتري محوري خود جامه عمل پوشاند. اين مقاله مفاهيمي از مديريت ارتباط با مشتري و يكي از عناصر آن - داده كاوي- را مورد بررسي قرار مي دهد.

مقدمه

در سالهاي اخير فرهنگ تجارت به پيشرفتهايي نايل گشته است. مطابق با آن روابط اقتصادي مشتريان به شيوه هاي بنيادي و اساسي در حال تغيير است. شركتها به منظور نظارت بر اينگـــونه تغييرات نيازمند ارايه راه حلها هستند. ظهور و پيدايش اينترنت در تغيير جهت مركز توجه بازاريابي نقش بسزايي داشته است. چنانچـــه اطلاعات بر خط (ON LINE) بيشتر در دسترس قرار گيرد موجب آگاهي و هوشياري بيشتر مشتريان مي گردد. آنها در جريان تمام آنچه ارايـــه و پيشنهاد مي شود قرار مي گيرند و تقاضاي بهترينها را دارند. براي از عهده برآمدن در چنين شرايطي بايد سيستم هايي كه بتواند به طور دقيق نسبت به مشتريان واكنش نشان دهد به كـــار رود. جمع آوري آمار مشتريان و داده هاي رفتاري آنها اين هدف اصلي و دقيق را ممكن مي سازد. اين نوع هدفگيري به يك برنامه ريزي عالي هنگام ايجاد يك رقابت سخت و به مشخص كردن مشتريان بالقوه هنگام عرضه محصولات جديد كمك مي كند.

داده كاوي

امروزه با حجم عظيمي از داده ها روبرو هستيم. براي استفاده از آنها به ابزارهاي كشف دانش نياز داريم. داده كاوي به عنوان يك توانايي پيشرفته در تحليل داده و كشف دانش مورد استفاده قرار مي گيرد. داده كاوي در علوم (ستاره شناسي،...)در تجارت (تبليغات، مديريت ارتباط با مشتري،...) در وب (موتورهاي جستجو،...) در مسايل دولتي (فعاليتهاي ضد تروريستي،...) كاربرد دارد. (١) عبارت داده كاوي شباهت به استخراج زغال سنگ و طلا دارد. داده كاوي نيز اطلاعات را كه در انبارهاي داده مدفون شده است، استخراج مي كند. (٢)

در واقع هـــدف از داده كاوي ايجاد مدل هايي براي تصميم گيري است. اين مدلها رفتارهاي آينده را براساس تحليلهاي گذشته پيش بيني مي كنند. به كاربردن داده كاوي به عنوان اهرمي براي آماده سازي داده ها و تكميل قابليتهاي انباره داده (DATA WAREHOUSE) ، بهترين موقعيت را براي به دست آوردن برتريهاي رقابتي ايجاد مي كند.

سيستم هاي بانك داده (BASE DATA) ، نقشي كليدي در سيستم هاي مديريت و انبار داده، بازي مي كنند. يك سيستم بانك داده، شامل فايل هاي بانك داده و سيستم هاي مديريت بانك داده است.(١)

اغلب تجارت ها به تصميم گيريهاي استراتژيك و يا اتخاذ خط مشي هاي جديد براي خدمت رساني بهتر به مشتريان نياز دارند. به عنوان مثال فروشگاهها آرايش مغازه خود را براي ايجاد ميل بيشتر به خريد مجدداً طراحي مي كنند و يا خطوط هواپيمايي تسهيلات خاصي را براي مشتريان جهت پروازهاي مكرر آنها در نظر مي گيرند. اين دو مثال به داده هايي در مورد رفتار مصرفي گذشته مشتريان براي تعيين الگوهايي به وسيله داده كاوي، نياز دارد. براساس اين الگوها تصميمـــات لازم اتخاذ مي شود. در واقع ابزار داده كـــاوي، داده را مي گيرد و يك تصوير از واقعيت به شكل مدل مي سازد، اين مدل روابط موجود در داده ها را شرح مي دهد. (٢)

از نظر فرايندي فعاليتهاي داده كاوي به سه طبقه بندي عمومي تقسيم مي شوند: (٦)

اكتشاف : فرايند جستجو در يك بانك داده براي يافتن الگوهاي پنهان، بدون داشتن يك فرضيه از پيش تعيين شده درباره اينكه اين الگو ممكن است چه باشد.

مانند تحليلهايي كه برحسب كالاهاي خريداري شده صورت مي گيرد، اينگونه تحليلهاي سبدي نشانگر مواردیست كه مشتري تمايل به خريد آنها دارند. اين اطلاعات مي تواند به بهبود موجودي، استراتژي طراحي، آرايش فروشگاه و تبليغات منجر گردد.

مدل پيش بيني : فرايندي كه الگوهاي كشف شـــده از بانك داده را مي گيرد و آنها را براي پيش بيني آينده به كار مي برد.

مانند پيش بيني فروش در خرده فروشي، الگوهاي كشف شده براي فروش به آنها كمك مي كند تا تصميماتي را در رابطه با موجودي اتخاذ كنند.

تحليلهاي دادگاهي : به فرايند به كارگيري الگوهاي استخراج شده براي يافتن عوامـــل داده اي نامعقول و متناقض مربوط مي شود.
مانند شناسايي و تشخيص كلاهبرداري در موسسات مالي. كلاهبرداري به ميزان زيادي پرهزينه و زيان آور است، بانكها مي توانند با تحليل دادوستدهاي جعلي گذشته الگوهايي را براي تشخيص و كشف كلاهبرداري به دست آورند.

مديريت ارتباط با مشتري

مديريت ارتباط با مشتري يك فرايند تجاري است كه تمام جوانب مشخصه هاي مشتري را آدرس دهي مي كند، دانش مشتري را بـــه وجود مي آورد، روابط را با مشتري شكل مي دهد و برداشت آنها را از محصولات يا خدمات سازمان ايجاد مي كند. مديريت ارتباط با مشتري توسط چهار عنصر از يك چارچوب ساده تعريف شده است: دانش، هدف، فروش و خدمت.(٢)

مديريت ارتباط با مشتري با درنظر گرفتن اينكه چه محصولات يا خدماتي، به چه مشترياني، در چه زماني و از طريق چه كانالي عرضه شود، بهبود را در پي خواهد داشت. اين مديريت از اجزاي مختلفي تشكيل شده است.

پيش از اينكه فرايند آن آغاز شود، شركت بايد اطلاعات مشتري را در اختيار داشته باشد. اين اطلاعات مي تواند از داده هاي داخلي مشتريان و يا از داده هاي منابع خارجي خريداري شده، به دست آيد. براي داده هاي داخلي منابع مختلفي وجود دارد مانند پـرسشنامه ها و بلاگ ها ، سوابق كارت اعتباري و....

منابع داده خارجي يا بانكهاي داده خريداري شده مانند آدرسها، شماره تلفن ها، پروفايل هاي بازديد از وب سايتها كليدي براي به دست آوردن دانش بيشتري از مشتري است.(۳)

بيشتر شركتها، بانكهاي داده اي عظيمي شامل داده هاي بازاريابي، منابع انساني و مالي را دارا هستند. بنابراين، سرمايه گذاري در زمينه انبار داده، يكي از اجزاي حياتي در استراتژي مديريت ارتباط با مشتري است. (٤)

پس از تهيه و تخصيص منابع داده، سيستم مديريت ارتباط با مشتري بايد با به كارگيري ابزارهايي مانند داده كاوي، داده ها را تجزيه و تحليل كند. اعم از اينكه شركت تكنيك هاي آماري سنتي را به كار مي برد يا يكي از ابزارهاي نرم افزاري مانند داده كاوي را، كارشناسان نياز به فهم داده هاي مشتري و روابط تجاري دارند. بنابراين، داشتـــن افرادي متخصص كه اين داده ها را با ابزارهاي مربوطه استخراج و به صورت اطلاعات درآورند، مهم است.

چرخه زندگي مشتري

واژه چرخه زندگي مشتري به مراحلي در ارتباط بين مشتري و تجارت بر مي گردد و آگاهي نسبت به آن موجب سودآوري مشتري مي شود. عموماً چهار مرحله در چرخه زندگي مشتري وجود دارد:

۱ - مشتريهاي بالقوه : افرادي كه هنوز مشتري نيستند ولي در هدف بازار قرار دارند؛

۲ - مشتريهايي كه عكس العمل نشان مي دهند: مشتريان بالقوه يا احتمالي كه به يك محصول يا خدمت علاقه و واكنش نشان مي دهند.

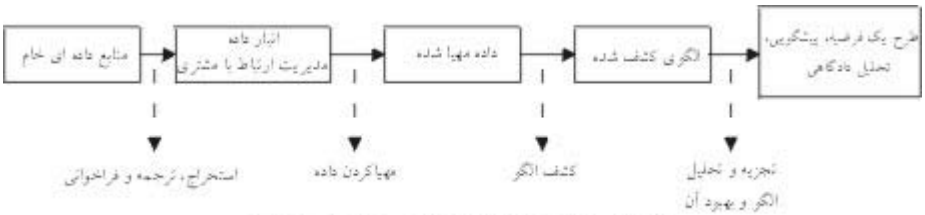
۳ - مشتريهاي بالفعل: افرادي كه در حال حاضر محصول يا خدمتي را به كار مي برند.

٤ - مشتريهاي سابق: اينگونه افراد مشتريان مناسبي نيستند چرا كه مدت زيادي در هدف فروش قرار ندارند و يا خريدشان را به سمت محصولات رقيب برده اند. (۲)

فرايند داده كاوي در مديريت ارتباط با مشتري

داده كاوي يكي از عناصر مديريت ارتباط با مشتري است و مي تواند به حركت شركتها به سمت مشتري محوري كمك كند.

فرايند داده كاوي در مديريت ارتباط با مشتري به صورت زير است . (شكل ۱)



شكل ١ – فرايند داده كاوي در مديريت ارتباط با مشتري

داده هاي خام از منابع مختلفي جمع آوري مي شوند و از طريق استخراج، ترجمه و فرايندهاي فراخواني به انبار داده اين مديريت وارد مي شوند. در بخش مهيـــاسازي داده، داده ها از انبار خارج شده و به صورت يك فرمت مناسب براي داده كاوي در مي آيند.

بخش كشف الگو شامل چهار لايه است:

١ - سوالهاي تجاري مانند توصيف مشتري،٢ - كاربردها مانند امتيازدهي، پيش گويي،٣ - روشها مانند سري هاي زماني، طبقه بندي،٤ - الگوريتم ها.
در اين بخش روشهاي داده كاوي با كاربرد مخصوص خود براي پاسخ به سوالهاي تجاري كه به ذهن مي رسند، الگوريتم هايي را استخراج مي كنند و از اين الگوريتم ها براي ساخت الگو استفاده مي شود.
در بخش تجزيه و تحليل الگو، الگوها به يك دانش مفيد و قابل استفاده تبديل مي شوند و پس از بهبود آنها، الگوهايي كه كارا محسوب مي شوند در يك سيستم اجرايي به كار گرفته خواهند شد.(١)

نتيجه گيري

رابطه مشتري با زمان تغيير مي كند و چنانچه تجارت و مشتري درباره يكديگر بيشتر بدانند اين رابطه تكامل و رشد مي يابد. چرخه زندگي مشتري چارچوب خوبي براي به كارگيري داده كاوي در مديريت ارتباط با مشتري فراهم مي كند. در بخش ورودي داده كاوي، چرخه زندگي مشتري مي گويد چه اطلاعاتي در دسترس است و در بخش خروجي آن، چرخه زندگي مي گويد چه چيزي احتمالاً جالب توجه است و چه تصميماتي بايد گرفته شود. داده كاوي مي تواند سودآوري مشتري هاي بالقوه را كه مي توانند به مشتريان بالفعل تبديل شوند، پيش بيني كند و اينكه تا چه مدت به صورت مشتريان وفادار خواهند ماند و چگونه احتمالاً ما را ترك خواهند كرد.

بعضي از مشتريان مرتباً مراجعاتشان را به شركتها براي كسب مزيتهايي كه طي رقابت ميان آنها به وجود مي آيد، تغيير مي دهند. در اين صورت شركتها مي توانند هدفشان را روي مشترياني متمركز كنند كه سودآوري بيشتري دارند.

بنابراين مي توان از طريق داده كاوي ارزش مشتريان را تعيين، رفتار آينده آنها را پيش بيني و تصميمات آگاهانه اي را در اين رابطه اتخاذ كرد.

منابع:

١ - .STATE UNIVERSITY, 2003 NONE YE, "THE HAND BOOK OF DATA MINING", ARIZONA

٢ - CHRIS RYGIELSKI, "DATA MINING TECHNIQUES FOR CUSTOMER RELATIONSHIP MANAGEMENT", TECHNOLOGY IN SOCIETY, 2002.
٣ - FREEMAN M., "THE CUSTOMER LIFECYCLES", INTELLIGENT ENTERPRISE, 1999.
٤ - HILL L., "CRM: EASIER SAID THAN DONE", INTELLIGENT ENTERPRISE, 1999.
٥ - CAP & IDC GEMINI, "FOUR ELEMENTS OF CUSTOMER RELATIONSHIP MANAGEMENT", CAP GEMINI WHITE PAPER.
٦ - INFORMATION DISCOVERY INC., "A CHARACTERIZATION OF DATA MINING TECHNOLOGIES AND PROCESSES", HTTP://WWW.DMREVIEW.COM/PORTAL

●Why Data Mining?

●What Is Data Mining?

●Potential Applications

●Data Mining Process

●Data Mining and Business Intelligence

●Interdisciplinary aspects of Data Mining

●Why Not Traditional Data Analysis?

●Multi- Dimensional view of Data Mining

●What Kinds of Data?

●Data Mining tasks

●Architecture: Typical Data Mining System

●History of Data Mining

●Conferences and Journals on Data Mining

●Where to Find References

●Data Preprocessing

●Data Warehousing and OLAP Technology

●Mining Frequent Patterns, Association and Correlations

●Classification and Predication (LDA, QDA, Decision trees, Neural Networks, Regression, Clustering, etc)

●Numerical Examples for Classification and Prediction using SPSS Clementine ●Applications and Trends in Data Mining

مرجع: کتاب داده کاوی - دکتر جمال شهرابی

داده کاوی

مقاله زیر را من از اینترنت گرفتم، متاسفانه منبع دقیق آن را یادم نیست. در هر صورت امیدوارم مفید باشد .

داده کاوی

داده کاوی فرایندی تحلیلی است که برای کاوش داده ها (معمولا حجم عظیمی از داده ها - در زمینه های کسب وکار و بازار) صورت می‌گیرد و یافته‌هابابه‌کارگیری الگوهایی،احراز اعتبار می‌شوند . هدف اصلی داده کاوی پیش بینی است. فرایند داده کاوی شامل سه مرحله می باشد : ۱. کاوش اولیه ۲. ساخت مدل یا شناسایی الگو با کمک احراز اعتبار/ تایید و ۳. بهره برداری.

مرحله ۱ : کاوش. معمولااین‌مرحله با آماده سازی داده ها صورت می گیرد که ممکن است شامل پاک سازی داده ها ،تبدیل داده هاوانتخاب زیرمجموعه‌هایی از رکوردهاباحجم‌عظیمی‌ازمتغ ?یرها(فیلدها) باشد . سپس با توجه‌به‌ماهیت‌مساله تحلیلی، این‌مرحله‌به‌مدل‌های پیش بینی ساده یا مدل‌های‌آماری‌وگرافیکی برای شناسایی متغیرهای مورد نظر و تعیین پیچیدگی مدل‌ها برای استفاده در مرحله بعدی نیاز دارد.

مرحله ۲:ساخت و احراز اعتبار مدل. این‌مرحله‌به بررسی‌مدل‌های مختلف و گزینش بهترین مدل با توجه به کارآیی پیش‌بینی آن می پردازد. شاید این مرحله ساده به نظر برسد، اما اینطورنیست.
تکنیک‌های‌متعددی‌برای‌رسیدن به‌این‌هدف توسعه یافتند.و " ارزیابی رقابتی مدل ها"نام گرفتند. بدین منظور مدل‌های مختلف برای مجموعه داده‌های یکسان‌به‌کارمی‌روندتاکار ?یی‌شان‌باهم مقایسه‌شود ،سپس مدلی که‌بهترین کارآیی راداشته باشد، انتخاب می‌شود.این‌تکنیک‌ها عبارتند از :
Stacking, Bagging,Boostingو.Meta-learning
مرحله ۳ : بهره برداری .آخرین‌مرحله‌مدلی‌راکه‌درمرح له قبل‌انتخاب‌شده است، در داده‌های‌جدیدبه کارمی‌گیردتا پیش‌بینی‌های‌خروجی‌های مورد انتظاررا تولید نماید.داده کاوی‌به‌عنوان‌ابزارمدیریت ?طلاعات‌برای‌تصمیم گیری،عمومیت‌یافته‌است .اخیرا،توسعه تکنیک های تحلیلی جدید در این زمینه مورد توجه قرار گرفته است (مثلا(Classification Trees ،اما هنوز داده کاوی مبتنی بر اصول آماری نظیر(Exploratory Data Analysis (EDA)می باشد.
بااین وجود تفاوت عمده ای بین داده کاوی و EDA وجوددارد.داده‌کاوی‌بیشترب ه‌برنامه‌های‌کاربردی گرایش دارد تا ماهیت اصلی پدیده. به عبارتی‌داده کاوی کمتر با شناسایی روابط بین متغیرها سروکار دارد.
مفاهیم اساسی در داده کاوی
:Baggingاین مفهوم برای ترکیب رده بندی های پیش بینی شده از چند مدل به کار می رود.فرض کنیدکه قصدداریدمدلی برای رده بندی پیش بینی بسازیدو مجموعه داده های مورد نظرتان کوچک است.شمامی توانید نمونه هایی(با جایگزینی) را از مجموعه داده ها انتخاب و برای نمونه های حاصل ازدرخت رده بندی (مثلا C&RT و(CHAIDاستفاده نمایید.به طورکلی برای نمونه های مختلف به درخت های متفاوتی خواهید رسید.سپس برای پیش بینی با کمک درخت های متفاوت به دست آمده از نمونه ها ،یک رای گیری ساده انجام دهید.رده بندی نهایی ، رده بندی ای‌خواهد بود که درخت های مختلف آنرا پیش بینی کرده اند.
:Boostingاین مفهوم برای تولید مدل‌های چندگانه (برای پیش بینی یا رده بندی)به کار می‌رود . Boostingنیزاز روش C&RT یا CHAID استفاده وترتیبی از classifier ها را تولید خواهد کرد.
: Meta-Learningاین مفهوم برای ترکیب پیش بینی‌های حاصل از چند مدل به کار می‌رود.و هنگامی که انواع مدل‌های موجود در پروژه خیلی متفاوت هستند، کاربرد دارد. فرض کنید که پروژه داده کاوی شما شاملTree classifier ها نظیرC&RT وCHAID ، تحلیل خطی و شبکه های عصبی است.هر یک از کامپیوترها،رده بندی هایی رابرای نمونه هاپیش بینی کرده اند.تجربه نشان می‌دهدکه ترکیب پیش بینی های چند روش دقیق تراز پیش بینی های هریک از روشهاست.پیش بینی های حاصل از چند classifierرا می توان به عنوان ورودی meta-linear مورد استفاده قرار داد meta-linear .پیش بینی هارا ترکیب می کند تا بهترین رده بندی پیش بینی شده حاصل شود.

متن مقاله :

امروزه **اطلاعات و فن آوري اطلاعات** آنچنان بر جنبه هاي مختلف زندگي سياسي، اجتماعي، فرهنگي و اقتصادي افراد و اجتماعات تاثير گذاشته است كه نمي توان از آن غفلت نمود. تاثير اطلاعات و فنآوري اطلاعات هم از جهت ايجاد فرصتهاي طلايي جديد و هم از نظر چالشهاي سازماني قابل مطالعه است . بدون شك جهاني شدن اقتصاد ، دسترسي به آن جوامع صنعتي و پيشرفته را به جامعه اطلاعاتي و اقتصادي آنها را از اقتصاد صنعتي به اقتصاد مبتني بر خدمات اطلاعاتي و دانش تبديل كرده است . فرآيند توليد و مديريت شركتهاي و سازمانها به شدت متحول شده اند و سازمانها و شركتهاي مجازي به وجود آمده اند . فن آوري اطلاعات بسياري از محدوديتهاي زماني و جغرافيايي را از بين بده است و افراد و سازمانها بدون توجه به مرزهاي جغرافيايي مي توانند با هم ارتباط برقرار كنند و به مبادله اطلاعات بپردازند. سيستمهاي اطلاعاتي ، ارتباط دوربرد و شبكه ها تحولات و دگرگونيهاي اساسي در توليد كالا و خدمات ، بازرگاني، بازريابي، فروش ، تبليغ و مديريت دسترسي به اطلاعات به وجود آورده اند. انقلاب اطلاعات كه از قرن بيستم شروع شده است ، تغييرات شگرفي را در نيروي كار به وجود آورده است و نيز حرفه ها و خدمات جديدي ظهور كرده اند كه به توليد ، توزيع يا كار با دانش و اطلاعات وابسته اند ، به طوري كه بيش از ٦٠ درصد توليد ناخالص ملي و حدود ٥٥ درصد نيريو كارگري آمريكا به اطلاعات و دانش وابسته است (صفحه ٣ همين كتاب).

در تعريف اقتصاد به عنوان يك علم توافق نظر كلي وجود ندارد. عده اي آن را مطالعه و بررسي توليد ، توزيع، و مصرف ثروت در يك جامعه انساني دانسته اند و گروهي ديگر آن را علم استفاده از منابع كمياب به منظور نيل به هدف رفاه اجتماعي براي جامعه بشري. در مورد اينكه اقتصاد به عنوان يك علم مطرح است يا هنر نيز بين صاحبنظران اختلاف وجود دارد . عده اي معتقدند هنگامي كه عملكردهاي اقتصادي با استفاده از رياضيات توجيه گردند. اقتصاد به عنوان يك علم مطرح است و بعضي ديگر معتقدند زماني كه از فلسفه و منطق براي توجيه مسايل اقتصادي استفاده كنند، اقتصاد به عنوان يك هنر مطرح خواهد بود. هر تعريفي را براي اقتصاد بپذيريم، مباحث اقتصادي مربوط به هر كالا يا خدمات به طور عمده شامل سرمابه گذاري، توليد، قيمت گذاري، بازاريابي، تبليغ، توزيع وفروش مصرف درآمد، شركتها، اشتغال و ساير مواردي است كه معمولاً در متون اقتصادي مطرح مي گردند. بر اين اساس اقتصاد اطلاعات مطالعه و بررسي توليد ، توزيع، بازاريابي ، قيمت گذاري، فروش، مصرف و كليه درآمدهايي است كه به طور مستقيم يا غير مستقيم از طريق توليد ، انتشار، فروش ، ذخيره ، پردازش، و دسترسي به اطلاعات حاصل مي شود. اطلاعات و دسترسي به آن چنان ارزش و اهميتي داردكه يكي از اركان اساسي توسعه و محور تصميم گيري در تمامي سطوح سياسي، اقتصادي، فرهنگي و اجتماعي است و به همين دليل هر سال سرمايه هاي هنگفتي براي تحقيق، توليد، انتشار، فراهم آوري، سازماندهي، ذخيره، بازيابي، اشاعه و دسترسي به آن صرف يم گردد و به عنوان كالايي با ارزش خريد و فروش مي گردد و بسياري از خدمات و صنايع پيچيده براي ذخيره، پردازش، انتقال و دسترسي به آن به وجو آمده اند و بسياري از كشروهاي پيشرفته بريا جذب اين بازار برنامه ريزي دقيق و سرمياه گذاري هنگفتي را انجام داده اند. اينترنت كه پديده اي جديد در فن آوري اطلاعات مي باشد،بر اساس بررسي دانشگاه تگزاس، حدود ٣٠٠ ميليارد دلار درآمد در سال ١٩٩٨ براي آمريكائيان داشته است . در حالي كه اقتصاد جهاني سالانه به طور متوسط ٣/٨ درصد و در آمريكا ٢/٨ درصد رشد دارد ، اقتصاد اينترنت به تنهايي در آمريكا به طور متوسط ١٧٤/٥ درصد از سال ١٩٩٥ تا ١٩٩٨ رشد داشته است و در اين

مدت كوتاه از صنايع قديمي و قدرتمندي جون انرژي (۲۲۳ ميليارد دلار درآمد سالانه)، اتومبيل (۳۵۰ ميليارد دلار) و ارتباطاطت راه دور (۲۷۰ ميليارد دلار) در آمريكا پيشي گرفته است . در حالي كه كشور آرژانتين از لحاظ درآمد ناخالص ملي در رتبه نوزدم در جهان مي باشد، درآمد حاصل از اينترنت در آمريكاه به تنهايي بيشتر از درآمد ناخالص ملي كشور آرژانتين مي باشد . اطلاعات، فنآوريهاي اطلاعات و ارتباطات به گونه اي توسعه و اهميت يافتند كه اين عصر را عصر اطلاعات، ارتباطات و رايانه خوانده اند و خدمات و مشاغل زيادي را به خود جذب كرده است . براي مثال : بر اساس مطالعه و پيش بيني دانشگاه تگزاس تا سال ۱۹۹۸ تنها در بخش فن آوري اينترنت۱/۲۰۰/۰۰۰ نفر شاغل بوده اند كه عمدتاً شغلهاي جديدي هستند . در انتشار و توزيع يك كتاب افرادي چون نويسنده،مترجم، ناشر، ويراستار، حروفچين، نمونه خوان، صفحه آرا، گرافيست ، ليتوگراف، چاپچي، صحاف، مراكز پخش ، كارگزاران و كتابفروشيها به طور مستقيم در ارتباطند. همچنين صنايع زيادي چون كاغذ ، چاپ، سخت افزارها و نرم افزارهاي رايانه اي نيز مستقيم و يا غير مستقيم به آن مربوط مي شوند . همچنين حرفه ها و مشاغل جديد ديگري چون فروشگاههاي الكترونيكي كتاب (مانند آمازون) ، توليد كنندگان و توزيع كنندگان كتابهاي الكترونيكي نيز جزئي از اقتصاد كتاب محسوب مي شوند كه خود بخشي از اقتصاد اطلاعات مي باشد . كشور آلمان در سال ۱۹۹٦، صرفنظر از ساير منابع اطلاعاتي ، تعداد ۷۱۵۱۵ عنوان كتاب چاپ اول منتشر شده است . اين بدان معني است كه نوسيندگان ، ناشران، و مشاغل و خدامت زيايدي در اين كشور به فعاليت مشغول بوده اند، به طوريكه تنها ارزش پشت جلد كتابها ۹ ميليارد دلار يا معادل ۹۰۰ ميليارد تومان ميباشد. آمار منتشر شده از سوي انجمن ناشران آمريكا حاكي از فروش ۲۳/۰۳۳/۳۰۰/۰۰۰ دلار كتاب در سال ۱۹۹۸ مي باشد . بر اساس آمار منتشر شده سال ۱۹۹۸ شركت ميكروسافت، كه تنها به توليد نرم افزار هاي رايانه اي مي پردازد، ۲۵۱۸۳ نفر را به طور تمام در اختيار داشته است و در آمد خالص اين شركت در همان سال ۱۱/۳٦۰/۰۰۰/۰۰۰ دلار بوده است كه نسبت به سال پيش ٥٤ درصد رشد داشته است. سفارش و تهيه ، سازماندهي ، حفاضت ، نگهداري، ذخيره ، پردازش و اشاعه اطلاعات، موجب تاسيس نهادها و شركتهايي چون كتابخانه، مراكز اطلاعات و اطلاع رساني ، بانكهاي اطلاعاتي، شركتهاي توليد انواع نرم افزار و سخت افزار و خدمات رايانه اي و بسياري ديگر از اين قبيل شده است . شبكه هاي اطلاعاتي عظيمي چون اينترنت به شدت در حلا دگرگون ساختن ارتباطات فرهنگي، اجتماعي ، اقتصادي و ... است و ارزش و اهميت اطلاعات و دسترسي به آن را دو چندان نموده است. مشاغل و خدمات جديدي را به وجود آورده و مرزهاي مكاني و زماني را در دسترسي به اطلاعات از بين برده است . كشورهاي پيشرفته با استفاده از فن آوري سنجش از راه دور و ماهواره اي به بسياري از اطلاعات اقتصادي و نظامي ساير كشروها دسترسي دارند و با استفاده از اين اطلاعات ، كه بسيار گران قيمت نيز هستند، روابط اقتصادي و نظامي خود را با ديگر كشورها تنظيم مي كنند.

با استفاده وسيع از فن آوريهاي جديد، اطلاعات به عنوان مواد خام به شكلهاي مختلف مورد داد و ستد قرار مي گيرد. مي توانيد دوره و شماره اي خاص از يك مجله را تصور كنيد كه گاهي به صورت چاپي و زماني به شكلهاي مختلف الكترونيكي (بر روي صفحه ارزان، صفحه فشرده يا سي.دي، دي.وي.دي. و اينترنت) مورد معامله قرار گيدر. مراكز چكيده نويسي و توليد پايگاههاي اطلاعاتي مقالات همين مجله را به صورت ديگري بسته بندي و به مشتريان خود عرضه مي نمايند . ميزبانها اطلاعات آن را به صورت پيوسته و از طريق شبكه هاي رايانه اي در اختيار كاربران خود قرار مي دهند. و به ازاي هر بار بازيابي مقاله اي خاص هزينه اي دريافت مي نمايند. هزينه دريافت اين

خدمات ممكن است بر ساسا ميزان ركورد بازيابي شده ، زمان اتصال و يا حتي سرعت انتقال اطلاعات باشد . متن كامل مقالات مجله مزبور زماني ديگر توسط مراكز خدمات تحويل مدرك بارها فروخته مي شود . با توجه به پيشرفتهاي اخير در توليد ، دخيره، بازيابي و انتقال اطلاعات ، بسياري از ناشران و مراكز خدمات تحويل مدرك نوعي ديگر از تجرات اطلاعات مربوط به همان مجله را رونق داده اند. تمامي خدمات مربوط به اشتراك، سفارش، دريافت، مسايل مالي و .. يك مجله از طريق اينترنت نيز صورت مي پذيرد و كاربر مي توان بهصورتهاي مختلف اطلاعات يك مجله را از طريق اينترنت و از كانالها و شركتهاي متفاوت دريافت نمايد: اطلاعاتي چون فهرست مندرجات، چكيده مقالات، خدمات آگاهي رساني جاري، اشاعه اطلاعات گزيده، دريافت متن كامل مجله ، دريافت متن كامل يك مقاله خاص و غيره. هر كدام از اين خدمات داراي نرخها و سيستمهاي قيمت گذاري جداگانه اي هستند.

به اين ترتيب نه تنها اطلاعات را به عنوان يكي از عوامل توليد در كنار عواملي جون كار، زمين و سرمايه قرار مي دهند كه در تمامي عرصه هاي فرهنگي ، اجتماعي ، اقتصادي و سياسي به عنوان عنصري اساسي و تاثير گذار رخ مي نمايد، بلكه خود موجب رونق صنايع اطلاعاتي، ارتباطي و ساير صنايع مرتبط با اطلاعات نيز شده است: به طوري كه بدون حضور اطلاعات بسياري از اين صنايع ارزش وجود خود را از دست خواهند داد.

به طور خلاصه، براي توليد، تهيه، سزاماندهي، ذخيره ، پردازش و دستري به اطلاعات فعاليتها ، خدمات و مشاغل زيادي به طور مستقيم يا غير مستقيم در ارتباطند كه بخشي از اقتصاد اطلاعات محسوب مي شوند. بريا مثال آن دسته از رايانه، سخت افزارها، نرم افزارها و مشاغلي كه بريا توليد و توزيع كتاب به كار رفته مي شوند، جزيي از اقتصاد اطلاعات محسوب مي شوند.

عليرغم آنكه اقتصاد اطلاعات در كشروهاي صنعتي و پيشرفته به خوبي از اهميت برخوردار است و اين كشورها برنامه ريزيهاي وسيعي را براي كسب درآمد از اين رهگذر شروع كرده اند، ولي در كشورهاي جهان سوم و در حال توسعه از جمله ايران اهميت آن به خوبي درك نشده است و يا هنوز زير ساختتهاي لازم براي توسعه اين بخش از اقتصاد وجود ندارد . در اينجا سعي مي شود در بخشهاي مختلف به طور مختصر به چشم اندازهاي اقتصادي اطلاعات اشاده گردد.

فن آوري اطلاعات

تكنولوژي اطلاع رساني / اطلاعاتي مجموعه ابزارها، ماشينها، دانش فني، روشها و مهارتهاي استفاده از آنها در توليد ، داد و گرفت ، پردازش، انباشت، بازيافت، جابه جايي، انتقال و مصرف اطلاعات است، از ساده ترين تا پيچيده ترين، و از ولتيترين تا پيشرفته ترين مراحل اطلاعاتي. با اين تعريف تمامي سخت افزارها مانند رايانه و لوازم جانبي آن و همچنين ابزارهاي ارتباط دوربرد، شبكه هاي اطلاع رساني، اينترنت، و نيز بسيار از نرم افزايها ديگر كه براي ذخيره، پردازش ، آماده سازي، بازيابي و مصرف اطلاعات به كار مي روند در اين رده قرار مي گيرند. همانگونه كه قبلاً نيز اشاره شد اين بخش از صنعت اطلاعات به سرعت در حال گسترش و دگرگوني است و ميليادها دلار درآند هر ساله عايد شركتهايي مي شود كه در اين بخش فعاليت دارند . هر چند توليد سخت افزايهاي اطلاع رساني در ايران به كندي در حال پيشرفت استو در مقايسه با كشورهاي پيشرفته فاصله زيادي را شاهد هستيم، ولي مديريت سخت افزار توليد نرم افزار بخشي از فن آوري اطلاعات است كه با برنامه ريزي دقيق مي توان از طريق آن درآمدهاي زيادي را نصيب شركتها و كشور كرد . توليد وفروش اين نرم افزارها مي توان در سطح ملي و بين المللي باشد. در سطح ملي نرم افزارهاي

خوب موجب استفاده بهنيه از سخت افزارهاي موجود براي دسترسي به اهداف سازماني و فردي نيز مي گردد، كه بريا آن هزينه هاي زيادي صرف شده است . براي مثال : توليد يك نرم افزار ذخير و بازيابي اطلاعات براي كتابخانه ها موجب صرفه جويي در هزينه، سرعت دسترسي به اطلاعات ، افزايش كيفي دسترسي به اطلاعات ، صرفه جويي در نيروي انساني، مديريت مناسبتر بر مجموعه و بسياري از مزاياي ديگر مي شود . كافي است بار ديگر اشاره كنيم كه در آمد خالص شركت مايكروسافت كه فقط به توليد نرم افزار مي پردازد در سال ۱۹۹۸ معادل ۱۱/۳٦۰/۰۰۰/۰۰۰ دلار بود ٥ است و نسبت به سال قبل ٥٤ درصد رشد داشته است. و يا انترنت در سال ۱۹۹۸، ۳۰۰/۰۰۰/۰۰۰/۰۰۰ دلار درآمد براي آمريكاييان داشته است و از سال ۱۹۹٥ دشدي معادل ۱۷٤/٥ درصد داشته است.

اشتغال

هر چند رشد صنعت اطلاعات موجب كاهش نياز به نيروي انساني در بسياري از بخشهاي توليد و خدماتي شده است ، ولي خود موجب ايجاد بسيار از شغلهاي جديد نيز شده است . براي مثال در اواخر دهه ۱۹۸۰ در كشور آفريقاي جنوبي در بخش اطلاع رساني ۲/۰۰۰/۰۰۰ نفر شاغل بودند . و يا در سال ۱۹۹۸ در ايالات متحده آمريكا ۱/۲۰۰/۰۰۰ نفر فقط در بخش اينترنت شاغل بودند. اگر در نظر بگيريم به طور متوسط حداقل دو نفر (با توجه به تيراژ در ايران) به طور مستقيم و غير مستقيم براي توليد ، توزيع و مصرف يك كتاب درگير باشند، افزايش عناوين و تيراژ كتاب و ساير منابع اطلاعاتي مي تواند در ايجاد شغل جديد بسيار موثر بشاد . بنابراين رشد توليد كتاب و ساير منابع اطلاعاتي نه تنها فعاليت فرهنگي و علمي است كه موجب ارتقاي سطح آگاهي و دانش مردم و كاهش بزهكاريهاي اجتماعي مي شود ، بلكه فعاليتي اقتصادي نيز محسوب مي شود . صنايع كاغذ سازي: فن آوري اطلاعات : تاليف و ترجمه ، حروفچيني ، نمونه خواني، ويراستاري ادبي و علمي، صفحه آرايي، فيلم و زينك، صحافي، چاپ، ناشران، كتابفروشي، توزيع عمده كتاب و سيار منابع اطلاعاتي چاپي و الكترونيكي : كتابخانه ها: توليد ، اطلاعات و اطلاع رساني است و جزئي از اقتصاد اطلاعات مي توانند محسوب شوند.

توليد توزيع و فروش اطلاعات

اطلاعات منبع اصلي تصميم گيري است و آنچنان اهميتي استراتژيك دارد كه آن را قدرت مي دانند. تصميم گيري درست و به موقع و برنامه ريزي كوتاه مدت و بلند مدت در مسائل اقتصادي، سياسي، فرهنگي، اجتماعي و ساير موارد مربوط به يك سازمان، شركت ، و يا در سطح ملي و بين المللي به اطلاعات مناسب و درست بستگي دارد و عدم دسترسي به آن خسارتهاي اقتصادي و مالي فروان و در بعضي مواقع جبران ناپذيري با به جاب خواهد گذاشت. توليد و انتشار اطلاعات علمي – فني ، اقتصادي، آماري و يا ساير اطلاعات مورد نياز و دسترسي به آن جزئي جدايي ناپذير از حيات اقتصادي يك كشور مي باشد. به طوري كه شركتهاي موفق وبين المللي براي رقابت در بازار و توليد محصول جديد و يا بهنيه سازي آن ميلياردها دلار صرف تحقيق و پژوهش مي نمايند. براي مثال صنايع داروسازي در آلمان سالانه بيش از ٥۰/۰۰۰/۰۰۰/۰۰۰ دلار براي تحقيق و پژوهش هزينه مي كنند . اطلاعات نوعي مواد خام است كه در كالاهاي ساخته شده موجود بوده و توليد كالا به وجود آن بستگي دارد. كالاهاي ساخته شده از بسياري جهات اطلاعات منجمد محسوب مي شوند. در جهان امروز با توجه به حجم اطلاعات منتشر شده ، بدون سازماندهي و پردازش اطلاعهات

دسترسي به بسياري از اطلاعات عملاً ناممكن است. به همين دليل سازماندهي، ذخيره، پردازش و توليد پايگاههاي اطلاعاتي و اشاعه آن بين متقاضيان اطلاعات، حرفه ها و مشاغل زيادي را به خود اختصاص داده است و از اين راه بسياري از سازمانها و شركتهاي انتفاعي و غير انتفاعي درآمدهاي هنگفتي را نصيب خود مي سازند. براي مثال كتابخانه ملي بريتانيا در سال ۱۹۹۸ روزانه بيش از ۱٤۰۰۰ مقاله به افراد متقاضي در سراسر دنيا ارسال كده است و از را ه درآمد سرشاري را نصيب خود كرده است . اين خدمات بخشي از فروش سالیانه كتابخانه ملي بريتانيا است و شامل فروش نرم افزار، پيگاههاي اطلاعاتي، خدمات مرجع و ساير خدمات جنبي ديگر نمي شود . نكته بسيار قابل تامل در فروش اطلاعات، در مقايسه باساير كالاها، اين است كه يك قلم از اطلاعات پس از يك بار فروش مجدداً نيز قابل فروش است . براي مثال يك مقاله ممكن است دهها بار فروخته شود چكيده آن به صورتي ديگر به فروش برسد در قالب فهرست مندرجات، خدمات آكاهي رساني جاري و اشاعه گزينشي اطلاعات به فروش برسد؛ و ... به كارگيري رايانه و شبكه هاي اطلاع رساني به خصوص اينترنت توليد ، سازماندهي و دسترسي به اطلاعات را متحول نموده است و مشاغل و شركتهاي جديدي را درگير نموده است . ناشران، مراكز خدمات تحويل مدرك و ميزبانان پايگاههاي اطلاعاتي از اين طريق به توليد ، بازريابي، فرورش و انتقال پاره هاي مختلف اطلاعاتي به مشتريان خود مي پردازند و افراد متقاضي اطلاعات حتي در منازل خود نيز مي توانند به اطلاعات مورد نياز دستري پيدا كنند. آنها براي هر جزء از اطلاعات و خدماتي كه ارائه مي نمايند هزينه دريافت مي كنند و مشتريان نيز با توجه به كيفيت و سرعت دسترسي به اطلاعات موجود در اين سيستمها به صورت روزانه و در بعضي مواقع حتي دقيقه اي روزآمد مي شوند و بدين ترتيب جاذبه هاي فراواني براي دسترسي به بسياري از اطلاعات به وجود آورده اند. با توجه به حجم زياد اطلاعات، واحد اطلاعات از يك كتاب ، مجله و به طور كلي از يك واحد كلي تر به جزئي از يك كتباب ، مجله و حتي قسمتي از يك مقاله و واحدي جزئي تر تبديل شده است و پايگاههاي اطلاعاتي زيادي بر اين اساس توليد و به بازار عرضه شده است . بسياري از محققان و جويندگان اطلاعات به جاي مراجعه به قفسه هاي كتابخانه و يا ساير واحدهاي نگهداري اطلاعات ابتدا به اين پايگاهها مراجعه و پس از انتخاب اوليه در اين پايگاهها بريا تهيه متن كامل مطلب مورد نياز به كتابخانه يا سيار واحدهاي نگهداري اطلاعات مراجعه مي كنند. بعضي از اين پايگاههاي اطلاعاتي استفاده كننده را مستقيماً به سايتي كه مقاله در آنجا موجود است متصل مي كنند و وي مي توان مطلب مورد نظر خود را مطالعه ، نسخه برداري و يا سفارش نمايد. بسته بندي مجدد و توليد اطلاعات سفارشي و ارائه سريع آن به متقاضيان، فعاليت ديگري است كه در حال توسعه است . توليد و توزيع اطلاعات به صورت الكترونيكي به خصوص از طريق اينترنت و نيز ارتباط مستقيم محققان با يكديگر و مبادله مستقيم و سريع اطلاعات و همچنين تشكيل گروههاي مباحثه و گروههاي خبري وساير روشهاي دسترسي به اطلاعات ، رفتار هاي اطلاع يابي محققان و دانشمندان را به شدت تحت تاثير قرار داده است . و ناشران را به چاش فراخوانده است. بسياري از شركتها و سازمانها معاملات تجاري، خريد و فروش و حتي پرداختهاي مالي مربوطه را از طريق اين نظامها و شبكه هاي اطلاع رساني انجام مي دهند . تجارت الكترونيكي چشم انداز جديدي است كه بسياري از شركتهاي معتبر را به سوي خود جلب نموده است.

به طور خلاصه در عصري كه آن را با نامهاي گوناگون عصر اطلاعات ، عصر كامپيوتر، عصر ارتباطات و عصر ماهواره مي نامند، يك يز به عنوان نقطه اشتراك تمامي اين نامگذاريها وجود دارد و آن اين است كه قطعات اطلاعات در پيكره اين ابزارها مبادله مي گردد. بدون حضور اطلاعات اين ابزارها

ارزش وجود خود را از دست خواهند داد . توليد و دسترسي به موقع به اطلاعات سياسي، اقتصادي ، علمي – فني و ساير اطلاعات مورد نياز موجب تصميم گيري خردمندانه است و قدرت محسوب مي شود به همين دليل سرمايه هاي هنگفتي براي توليد، ذخيره، پردازش و دسترسي به اطلاعات صرف مي گردد و صنايع اطلاعاتي و اقتصاد اطلاعات با نرخ رشدي شگفت آور در حال دگرگون كردن موازنه هاي اقتصادي است . صنايع پيچيده و ظريف بيش از پيش به اطلاعات و فن آوري اطلاعات وابسته است. اطلاعات نوعي مواد خام است كه در كالاهاي ساخته شده موجود بوده و توليد كالا به وجود آن بستگي دارد . كالاهاي ساخته شده از بسياري جهات اطلاعات منجمد محسوب مي شوند.

حميد محسني
عضو هيات علمي مركز اطلاعات و مدارك علمي وزارت كشاورزي

موسسه پژوهشی داده پردازان گیتا - دبیرخانه دائمی کنفرانس داده کاوی

به شرکت کنندگان در هر دوره (مقدماتی ، متوسط ، پیشرفته) که در تمامی کارگاههای آموزشی آن دوره شرکت نموده اند گواهینامه معتبر اعطا می گردد .

به شرکت کنندگانی که دارای حداقل یکی از شرایط زیر باشند حداکثر ۱۵% تخفیف تعلق می گیرد .

- شرکت کنندگان در هر سه دوره مقدماتی ، متوسط و پیشرفته

- ارائه کارت دانشجویی

- شرکت گروهی بیش از ۵ نفر

دوره مقدماتی (E)

عنوان کارگاه	کد کارگاه	پیش نیاز
مروری بر دانش داده کاوی (1) Data Mining Concepts (1)	E1	ندارد
مروری بر دانش داده کاوی (2) Data Mining Concepts (2)	E2	E1
فرایند و معماری داده کاوی Architecture and Data Mining Process	E3	ندارد
زمینه کاری و فعالیت داده کاوی Data Mining Tasks	E4	ندارد
الگوریتم ها و تکنیک های داده کاوی Data Mining Techniques and Algorithms	E5	ندارد
پروژه ها و کاربرد های داده کاوی Data Mining Projects and Applications	E6	ندارد

مدرس : جناب آقای دکتر جمال شهرابی

هزینه ثبت نام کل دوره : ۵۰۰۰۰۰ ریال

هزینه به ازای هر کارگاه : ۱۰۰۰۰۰ ریال

تاریخ برگزاری : ۲۲ تیرماه ۱۳۸۷

مکان : دانشگاه صنعتی امیرکبیر

دوره متوسط (I)

عنوان کارگاه	کد کارگاه	پیش نیاز
قوانین وابستگی Data Mining Task: Association Rules	I1	E
خوشه بندی Data Mining Task: Clustering	I2	E
کلاس بندی Data Mining Task: Classification	I3	E
مدل های پیش بینی پیشرفته Advanced Data Mining Task: Prediction Models	I4	E
داده کاوی زمان محور / مکان محور Data Mining Task: Spatial and Temporal Data Mining	I5	E
متن کاوی / وب کاوی Data Mining Task: Text Mining / Web Mining	I6	E

مدرس : جناب آقای دکتر جمال شهرابی

هزینه ثبت نام : ۸۰۰۰۰۰ ریال

هزینه به ازای هر کارگاه : ۱۵۰۰۰۰ ریال

تاریخ برگزاری : ۲۳ تیرماه ۱۳۸۷

مکان : دانشگاه صنعتی امیرکبیر

دوره پیشرفته (A)

عنوان کارگاه	کد کارگاه	پیش نیاز
مروری بر داده کاوی در SQL Server۲۰۰۵ Over View of SQL Server۲۰۰۵ Data Mining	A۱	E
تحلیل سبد بازار در SQL Server۲۰۰۵(DMX) با استفاده از قواعد وابستگی Market Basket DMX Tutorial with Association Rules in SQL Server۲۰۰۵	A۲	E,I۱,A۱
بخش بندی بازار در SQL Server۲۰۰۵(DMX) با استفاده از خوشه بندی Segmentation DMX with Clustering in SQL Server۲۰۰۵	A۳	E,I۲,A۱
کلاس بندی و پیش بینی در SQL Server۲۰۰۵(DMX) با استفاده از درخت تصمیم Classification and Prediction DMX with Decision Tree in SQL Server۲۰۰۵	A۴	E,I۳,A۱
ارزیابی صحت مدل ها با استفاده از نمودار صحت Evaluation with View Mining Accuracy Charts	A۵	E,A۱

مدرس : سرکار خانم مهندس ونوس شکورنیاز

هزینه ثبت نام : ۱۰۰۰۰۰۰ ریال

تاریخ برگزاری : ۲٤ تیرماه ۱۳۸۷

مکان : دانشگاه صنعتی امیرکبیر

Data Mining Tasks

- Classification/partitioning
- Clustering
- Association
- Segmentation
- Regression
- Advanced prediction modeling
- Temporal data mining
- Spatial data mining
- Time series forecasting
- Deviation and outlier detection
- Explorative and visual data mining
- Web mining
- Text Mining
- Mining semi-structured data
- Content mining and pattern mining
- Multimedia mining (audio/video)
- Explorative and visual data mining
- Others

Data Mining Algorithms

- Clustering algorithms
- Genetic algorithms and categorization techniques
- Fuzzy logic and rough sets

Conference Topics

Data Mining Process

- Data preparation techniques
- Data reduction methods
- Data cleaning and preparation
- Feature selection and transformation
- Sampling and rebalancing
- Missing value imputation
- Model selection/assessment and comparison
- Model comparison
- Model interpretation
- Others

Data Mining Applications

- Engineering
- System and Manufacturing
- Industry and government
- System planning and management
- Urban planning and management
- Logistics/Traffic management
- Science and technology
- Education
- Business/Industrial

- Artificial neural networks
- Decision trees/rule learners
- Statistical methods
- Case based reasoning
- Link and sequence analysis
- Others

Data Mining Integration

- Mining large scale data
- Multidimensional data
- Distributed and grid based data mining
- Data visualization
- Knowledge Discovery in Databases (KDD)
- Data and knowledge representation
- Data warehousing
- OLAP integration
- Others

- Marketing
- Finance and financial services
- Insurance
- Social science
- Military/Security
- Bioinformatics/Medicine
- Biological sciences
- Risk analysis
- Emergency planning services
- Health, safety and environment (HSE)
- Others

حالت اول: فرض کنید یک بزه‌کار اقتصادی بخواهد یک میلیون دلار را از طریق واردات کالا پول‌شویی کند. ابتدا لازم است به عنوان یک وارد کننده محلی، با یک صادرکننده خارجی (شایدخودش) تبانی کند. سپس کارهای زیر را انجام دهد:
۱- صادرکننده خارجی ده هزار تیغ را بازای هرعدد یک دهم سنت می‌خرد. (جمعا ۱۰۰۰ دلار)
۲- صادرکننده خارجی ده هزار تیغ را به یک صادرکننده محلی به قیمت هرعدد ۱۰۰ دلار می‌فروشد. (جمعا ۱ میلیون دلار).
۳- واردکننده محلی ده هزار تیغ را به قیمت واقعی هزار دلار دریافت می‌کند و یک میلیون دلار به صادرکننده خارجی می‌پردازد.
٤- نتیجه: واردکننده محلی یک میلیون دلار را به یک کشور خارجی با هزینه هزار دلار انتقال (شست و شو) داده است.
عمل فوق با اضافه صورت حساب کردن در واردات کالا صورت می‌گیرد.

حالت دوم: فرض کنید یک بزه‌کار اقتصادی بخواهد یک میلیون دلار را ازطریق صادرات کالا پول‌شویی کند. ابتدا لازم است به عنوان یک صادر کننده محلی، با یک وارد کننده خارجی تبانی‌کند. سپس کارهای زیر را انجام دهد:
۱- بزه‌کار محلی، با یک میلیون دلار، ۲۰۰ ساعت طلای تجملی به قیمت هریک پنج هزار دلار با پول نقد می‌خرد. (جمعا یک میلیون دلار).
۲- صادرکننده محلی، ۲۰۰ ساعت طلا را به یک وارد کننده خارجی به قیمت هر عدد ٥ دلار می‌فروشد. (جمعا یک هزار دلار).
۳- واردکننده خارجی ، ۲۰۰ ساعت طلا را دریافت کرده و یک هزار دلار برای صادر کننده محلی صورت حساب می‌کند.
٤- صادر کننده خارجی، ساعت‌های طلا را در بازار به قیمت هریک پنج هزار دلار می‌فروشد. (جمعا یک میلیون دلار).
نتیجه: صادر کننده محلی یک میلیون دلار به یک کشور خارجی با هزینه ۱۰۰۰ دلار انتقال (شست و شو) داده است.
عمل فوق با کم صورت حساب کردن در صادرات کالا صورت می‌گیرد.

استفاده از تجارت جهانی برای انتقال پول سیاه از یک کشور به کشور دیگر، یکی از روش‌های قدیمی برای فرار از حسابرسی دولتی است. این کار از طریق اضافه صورت حساب کردن واردات یا کم صورت حساب کردن صادرات انجام می‌شود. البته برعکس هر دو عمل فوق نیز امکان دارد. سازمان‌های اقتصادی و اطلاعاتی معمولا کارهایی برای کشف پول‌شویی از طریق درهای جلو (موسسات مالی و حسابرسی) انجام داده و از درهای پشت ساختمان (تجارت بین الملل) مقداری غافل می‌مانند. البته برعکس حالت فوق نیز در ایران رایج است.
زیاد صورت حساب شدن واردات سه خلاف را ممکن است در بر داشته باشد:
۱- خلاف گمرکی.
۲- فرار از مالیات.
۳- پول‌شویی.
این پول‌ها ممکن است در اختیار سازمانی مثل القاعده قرار گیرد. عملیات پول‌شویی از این راه‌ها و انتخاب بهترین راه با مطالعه شرایط محیطی هرکشور صورت می‌گیرد. بعنوان مثال در آمریکا چون کنترل کمتری روی صادرات کالا می‌شود، بهترین راه، از طریق صادرات کالا به دیگر کشورها است .

روش داده‌کاوی (Mining Data):
تحلیل اطلاعات واردات و صادرات کالا در وزارت بازرگانی آمریکا، ممیزی مالیاتی و بانک اطلاعاتی کالاهای تجاری نشان داده است که با روش‌های داده‌کاوی می‌توان به سرنخ‌هایی دست پیدا کرد. با داشتن قیمت تقریبی یک واحد کالا در زمان مشخص می‌توان کالاهای صادر و وارد شده را بررسی و انحراف‌های غیرعادی و زیاد را نشان کرده و ممیزی نمود. این کار برای ۱٦۳۹۰ کالای وارداتی و ۸٥٦۸ کالای صادراتی در سال ۲۰۰۱ و برای ۲۳۰ کشور که با آمریکا مراوده بازرگانی داشتند انجام شد. تمام واردات و صادرات با حد بالا و پائین مقایسه و ثبت‌گردید. مقادیر دلاری و تعداد موارد مشکوک برای هرکشور تجمیع شد. کل پول منتقل شده به خارج از آمریکا در سال ۲۰۰۱ مبلغ ۱٥٦ میلیارد دلار بود.
جدول ۱ نمونه‌ای از موارد فوق است.

کالا	کشور	قیمت واحد
دستمال کاغذی خشکبار قیچی تیغ	چین بلژیک ژاپن انگلیس	٤١٢١ دلار هر هزار گرم ٢٤٢٦ دلار هر هزار گرم ٤٨٩٦ دلار هر عدد ١١٣ دلار هر عدد
کشورهای مظنون به وجود نیروهای القاعده در آن‌ها		
حوله پنبه ای آیینه تیغ تلمبه دستی میل لنگ	پاکستان اندونزی مصر مالزی عربستان	١٥٤ دلار هر عدد ١٦٥ دلار هرسانتی مترمربع ٢٣ دلار هر واحد ٥٠٠٠ دلار هر واحد ١٥٢٠٠ دلار هر واحد

جدول ١ قیمت‌های غیرعادی واردات آمریکا

مقدار محاسبه شده که احتمال دارد توسط القاعده جابجا شده باشد حدود ٢٧ / ٤ میلیارد دلار برای ٢٥ کشور در فهرست مظنونین می‌باشد. برای پنج کشور بالای فهرست حدود ٦٥ / ٣ میلیارد دلار است. با توجه به مقدار کل ١٥٦ میلیارد دلار پول‌شویی که از این راه محاسبه شده و مقدار ٢٧ / ٤ میلیارد دلار که القاعده مظنون به پول‌شویی است، حدود ٣ درصد برای القاعده و ٩٧ درصد برای قالتاق‌ها و بزه‌کاران اقتصادی آمریکا و جهان می‌شود. جدول ٢ نمونه‌ایی ازصادرات غیرعادی از آمریکا به دیگر کشورها می‌باشد.

کالا	کشور	قیمت واحد
الماس تزئینی بالابر بلدوزر خودکششی پروژکتور ویدیو موشک و سکوی پرتاب	هند جامائیکا کلمبیا برزیل اسرائیل	١٣ دلار هر قیراط ٣٨٤ دلار هر دستگاه ١٧٤٢ دلار هر دستگاه ٣٤ دلار هر دستگاه ٥٢ دلار هر واحد
کشورهای مظنون به وجود نیروهای القاعده در آن‌ها		
مانتیور رنگی ویدیو مانیتور رنگی ویدیو کفش ورزشی ایزوتوپ رادیواکتیو	اندونزی پاکستان اردن مصر	٢٣ دلار هر دستگاه ٢٢ دلار هر دستگاه ٤ دهم دلار هر جفت ١ صدم دلار هر واحد

جدول ٢ قیمت‌های غیرعادی صادرات آمریکا

نتیجه‌گیری : با توجه به روش‌ها و امکان داده کاوی، می‌توان از اطلاعات بانک‌های عامل و بانک مرکزی ، وزارت بازرگانی ، گمرکات کشور ،پایانه‌های حمل و نقل کالا، وزارت صنایع و معادن و دیگر سازمان‌های ذیربط استفاده نموده و موارد مشکوک را ممیزی کرد.
پیش نیاز فوق وجود راه‌ها و اطلاعات زیاد و کافی در سازمان های مربوطه است. اگر سامانه‌هایی در سازمان‌های مربوطه وجود داشته باشند که بتواند اطلاعات را بصورت "آن لاین" و فوری در اختیار بگذارند، باتوجه به اعلام بازرگانان در مبادی حمل و نقل، صدور و ورود کالا، می‌توان قبل از اقدام به صادر یا وارد کردن کالا، آن را ممیزی و کشف نموده و مانع از پول‌شویی و فساد اقتصادی گردید. کارهای آماری و اطلاعاتی از این نوع یک علم تقریبی است، اما با امکانات رایانه‌ایی و روش‌های داده‌کاوی (Data Mining)، می‌توان کیفیت و دقت آنرا بیشتر کرد.

داده کاوی (DataMining)

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

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

مدیریت ذخیره سازی و دستیابی اطلاعات

داده های اطلاعاتی (Data)به عنوان یکی از منابع حیاتی سازمان شناخته می شود و بسیاری از سازمان ها با اطلاعات و دانش سازمانی خود مانند سایر دارایی های ارزشمندشان برخورد می کنند. نکته :داده اطلاعاتی (Data) به اطلاعات خام سازمان اتلاق می‌شود و اطلاعات (Information) به داده‌های پردازش شده .همچنین داده های پردازش شده پس از طبقه بندی و آنالیز به دانش سازمان (Knowledge)تبدیل می گردند.

کاوش‌های ماشینی در داده‌ها یا داده‌کاوی (Data mining) را باید یکی از سامانه‌های هوشمند (Intelligent systems) دانست. سامانه‌های هوشمند زیر شاخه‌ایست بزرگ و پرکاربرد از یادگیری ماشینی که خود زمینه‌ای‌ست در هوش مصنوعی. زمینه علمی جدید و پهناور یادگیری ماشینی (که "کاوش‌های ماشینی در داده‌ها" بخشی ست بزرگ از زیر شاخه سامانه‌های هوشمند آن ست)، به واقع

همان امتداد و استمرار دانش کهن و همه جا گیر آمار است در جهت ماشینی کردن یادگیری، تعلّم، و سرانجام، دانش.

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

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

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

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

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

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

تحلیلهای دادگاهی : به فرایند به کارگیری الگوهای استخراج شده برای یافتن عوامـــل داده ای نامعقول و متناقض مربوط می شود.

مانند شناسایی و تشخیص کلاهبرداری در موسسات مالی. کلاهبرداری به میزان زیادی پرهزینه و زیان آور است، بانکها می توانند با تحلیل دادوستدهای جعلی گذشته الگوهایی را برای تشخیص و کشف کلاهبرداری به دست آورند.

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

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

ان :داده كاوي چيست ؟
نویسنده/نویسندگان:ترجمه و تالیف : مهندس عذرا قبادي
چکیده :جامعه مبتنی بر اطلاعات را می توان به عنوان جامعه ای تعریف نمود که بخش غالب اجتماع به جای کارهای فیزیکی در گیرکارهاي فکری هستند .در چنین جامعه ای بیشترین توجه به فعالیتهای اطلاعاتی از قبیل : فراهم آوری، پردازش، تولید، ثبت، انتقال، اشاعه و مدیریت اطلاعات مبذول می گردد وبیشترین هزینه ها صرف فرایندهای اطلاعاتی می شود .(cawkell,1987) با گسترش سيستمهاي پايگاهي و حجم بالاي داده ها ي ذخيره شده در اين سيستم ها ، به ابزاري نيازست تا بتوان اين داده ها راپردازش كرد و اطلاعات حاصل از آن را در اختيار كاربران قرار داد.معمولا"كاربران پس از طرح فرضيه اي بر اساس گزارشات مشاهده شده به اثبات يا رد آن مي پردازند ، در حالي كه امروزه به روشهايي نيازداريم كه به اصطلاح به كشف دانش (Knowledge Discovery)بپردازند يعني روشهائي كه با كمترين دخالت كاربر و به صورت خ

كلمات كليدی :داده كاوي، توليد، اشاعه و مديريت اطلاعات و فرآيندهاي اطلاعاتي

داده كاوي چيست ؟

ترجمه و تاليف : مهندس عذرا قبادي

رئيس اداره نگهداري وپشتيباني سيستم ها

جامعه مبتنی بر اطلاعات را می توان به عنوان جامعه ای تعریف نمود که بخش غالب اجتماع به جای کارهای فیزیکی در گیرکارهاي فکری هستند. در چنین جامعه ای بیشترین توجه به فعالیتهای اطلاعاتی از قبیل : فراهم آوری، پردازش، تولید، ثبت، انتقال، اشاعه و مدیریت اطلاعات مبذول می گردد وبیشترین هزینه ها صرف فرایندهای اطلاعاتی می شود (cawkell,1987).

با گسترش سيستمهاي پايگاهي و حجم بالاي داده ها ي ذخيره شده در اين سيستم ها ، به ابزاري نيازاست تا بتوان اين داده ها راپردازش كرد و اطلاعات حاصل از آن را در اختيار كاربران قرار داد.معمولا"كاربران پس از طرح فرضيه اي بر اساس گزارشات مشاهده شده به اثبات يا رد آن مي پردازند ، در حالي كه امروزه به روشهايي نيازداريم كه به اصطلاح به كشف دانش(Knowledge Discovery) بپردازند

يعني روشهائي كه با كمترين دخالت كاربر و به صورت خودكار الگوها و رابطه هاي منطقي را بيان نمايند.

يكي از روشهاي بسيار مهمي كه با آن مي توان الگوهاي مفيدي را در ميان داده ها تشخيص داد،داده كاوي است ،اين روش كه با حداقل دخالت كاربران همراه است اطلاعاتي را در اختيار آنها وتحليل گران قرار ميدهد تا براساس آنها تصميمات مهم و حياتي در سازمانشان اتخاذ نمايند .

بايد توجه داشت كه اصطلاح داده كاوي زماني به كار برده مي شود كه با حجم بزرگي از داده ها ، در حد مگا يا ترابايت ، مواجه باشيم . در تمامي منابع داده كاوي بر اين مطلب تاكيد شده است . هر چه حجم داده ها بيشتر و روابط ميان آنها پيچيده تر باشد دسترسي به اطلاعات نهفته در ميان داده ها مشكلتر مي شود و نقش داده كاوي به عنوان يكي از روشهاي كشف دانش ،آشكارتر مي گردد.

داده كاوي از چندين رشته علمي بطور همزمان بهره ميبرد نظير : تكنولوژي پايگاه داده،هوش مصنوعي ،شبكه هاي عصبي،آمار،سيستم هاي مبتني بر دانش،بازيابي اطلاعات وغيره .]1[كه براي پرهيزاز اطاله كلام مي توان آن به لحاظ تاريخي به اختصار به مراحل زير تقسيم كرد:

مرحله اوليه: گردآوري وايجاد پايگاه اطلاعاتي(تا دهه ١٩٦٠)

مرحله دوم : نظامهاي مديريتي مبني بر پايگاه اطلاعاتي(دهه١٩٧٠ واوايل دهه ١٩٨٠)

مرحله سوم : نظامهاي پايگاه اطلاعاتي پيشرفته (اواسط دهه ١٩٨٠ تازمان حاضر)

مرحله چهارم : انبارش اطلاعات وداده كاوي (اواخر دهه ١٩٨٠ تا به امروز)

مرحله پنجم : نظام پايگاه اطلاعاتي مبني برشبكه (دهه ١٩٩٠ تا كنون)

مرحله ششم : نسل نوين نظامهاي اطلاعاتي يكپارچه شده(از ٢٠٠٠ به بعد)

بدين ترتيب فعاليتي كه از دهه ١٩٦٠ شروع شده بود در دهه ١٩٩٠ گامهاي بلندي برداشت وانتظار ميرود در اين قرن به رشد و بالندگي خود ادامه دهد.

تعريفي از داده كاوي

بطور كلي، داده كاوي (كه گاهي اوقات اكتشاف اطلاعات يا دانش ناميده ميشود)عبارت از فرآيندي است كه

از چشم اندازهاي مختلف به تحليل داده ها مي پردازد و جمعبندي آنها را در قالب اطلاعات مفيدي ارائه ميكند . اين اطلاعات را ميتوان براي افزايش در آمد ، كاهش هزينه ها يا هردوبه كاربرد. نرم افزار داده كاوي يكي از ابزارهاي تحليل اطلاعات است .اين نرم افزار به كاربران امكان مي دهد اطلاعات را از ابعاد وزواياي بسيار متفاوت تحليل و طبقه بندي كنند وروابطي را كه در آن ها شناسائي نموده اند به اجمال بيان نمايند.

به لحاظ فني، داده كاوي عبارت از فرآيندي است كه در ميان حوزه هاي گوناگون بانكهاي اطلاعاتي ارتباطي

بزرگ، همبستگي ها يا الگوهايي را پيدا مي كند. البته اين ويژگي به معناي يكسان دانستن داده كاوي وآناليزآماري نيست كه درجدول زير اين تفاوتهاآورده شده است :

آناليز آماری	*داده كاوي*
آمارشناسان همیشه با یک فرضیه شروع به کار میکنند. آمارشناسان باید رابطه هایی را ایجاد کنند که به	به فرضیه احتیاجی ندارد. الگوریتمهای داده کاوي در ابزارها بطور اتوماتیک

فرضیه آنها مربوط شود.	روابط را ایجاد میکنند.
آنها از داده های عددی استفاده میکنند.	ابزارهای داده کاوي از انواع مختلف داده و نه فقط عددی میتوانند استفاده کنند.
آنها میتوانند داده های نابجا و نادرست را در طول آنالیز تشخیص دهند.	داده کاوي به داده های صحیح و درست طبقه بندی شده بستگی دارد.
آنها میتوانند نتایج کار خود را تفسیر کنند و برای مدیران بیان کنند.	نتایج داده کاوي آسان نیست و همچنان به متخصصان آمار برای تحلیل آنها و بیان آنها به مدیران نیاز است.

پنج ویژگي مهم داده کاوي عبارت است از :

- استخراج ،دگرگوني وبارنمودن داده هاي تراکنشي برروي سیستم انبار داده ها .
- ذخیره ومدیریت داده ها در سیستم بانك اطلاعات چند بعدي.
- فراهم آوردن امکان دسترسي تحلیل گران تجاري ومتخصصان تحلیل اطلاعات به داده ها .
- تحلیل داده ها با استفاده از نرم افزارکاربردي .
- معرفي نمودن ، در یك قالب بندي سودمند ، همانند گراف یا جدول

داده کاوي به چه کار مي آید؟

امروزه در درجه اول شرکتها ازداده کاوي استفاده مي کنند.(با توجه بسیار زیاد به مصرف کننده ، خرده فروشي،مالي ، ارتباط،وسازمانهاي بازاریابي). داده کاوي این شرکتهارا قادر مي سازد که رابطه عوامل "دروني" (مانند قیمت ، موقع یابي فرآورده ،یا مهارت هاي کارمندان) ، را با عوامل "خارجي" (مانند شاخص هاي اقتصادي ، رقابت وآمارگیري جمعیتي مشتري) مشخص کنند؛ داده کاوي شرکت ها را قادر مي سازد اثر گذاري بر مشتري ، رضایتمندي مشتري و منافع شرکت را تعیین کنند.بالاخره ، شرکتها را قادر مي سازد که فشرده اطلاعات را براي دیدن داده هاي معاملاتي دقیق"حفاري" نمایند.

برخي از کاربردهاي داده کاوي در محیطهاي واقعي عبارتند از :

۱. خرده فروشي : از کاربردهاي کلاسیک داده کاوي است که مي توان به موارد زیر اشاره کرد :
   - تعیین الگوهاي خرید مشتریان
   - تجزیه و تحلیل سبد خرید بازار
   - پیشگویي میزان خرید مشتریان از طریق فروش الکترونیکي
۲. بانکداري :
   - پیش بیني الگوهاي کلاهبرداري از طریق کارتهاي اعتباري

- تشخيص مشتريان ثابت
- تعيين ميزان استفاده از كارتهاي اعتباري بر اساس گروههاي اجتماعي

۳. بيمه :

- تجزيه و تحليل دعاوي
- پيشگويي ميزان خريد بيمه نامه هاي جديد توسط مشتريان

٤. پزشكي :

- تعيين نوع رفتار با بيماران و پيشگويي ميزان موفقيت اعمال جراحي
- تعيين ميزان موفقيت روشهاي درماني در برخورد با بيماريهاي صعب العلاج]۲[

نتيجه اينكه :

بسياري از سازمانها بر معادني از طلا تكيه زده اند .اين گنجينه گرانبها در شركتهاي بيمه همان داده هاي جمع آوري شده از بيمه گذاران،بيمه شدگان ، زيانديدگان ، مقصران حادثه وانواع بيمه هاي فروخته شده است كه مي بايد با بهره گيري از تكنولوژيهاي جديد وابزارهاي خودكاري كه بصورت هوشمند آنهارا تجزيه وتحليل مي كنند ،گردآوري وپردازش شده وبه دانش تبديل وبه كار گرفته شوند

انبار داده ها

يک شنبه,۲۹ مهر ۱۳۸۳) تعداد دفعات خوانده شده:۲۸۸۳(

از اواسط سال هاي ۱۹۸۰ نياز به انبار داده ها به وجود آمد و دريافتند كه سيستم هاي اطلاعاتي بايد به صورت سيستم هاي عملياتي واطلاعاتي مشخص شوند. سيستم هاي عملياتي از فعاليت هاي روزانه كسب و كار پشتيباني مي نمايند و براي پاسخگويي سريع به ارتباطات از پيش تعريف شده مناسب هستند. داده هاي عملياتي ارائه بي درنگ و فعلي وضعيت كسب و كار مي باشند. اما سيستم هاي اطلاعاتي براي مديريت و كنترل كسب و كار به كار مي روند .

اين سيستم ها از تجزيه و تحليل داده ها براي اتخاذ تصميم درباره عملكرد آني و آتي سازمان پشتيباني مي كنند و براي درخواست هاي موردي، پيچيده و به طور كلي فقط خواندني طراحي شده اند. داده هاي اطلاعاتي تاريخي هستند، به عبارتي بيانگر ديدگاه ثابتي از كسب و كار در يك دوره زماني مي باشند .

ويژگي هاي اصلي داده هاي انبار داده ها

داده هاي موجود در انبار داده ها از سيستم هاي عملياتي متنوع (نظير سيستم هاي پايگاه داده ها) و منافع داده اي خارجي (نظير پايگاه داده هاي آماري و (WWWيكپارچه مي شوند. تفاوت هاي ساختاري و معنايي داده ها بايد پيش از يكپارچه سازي انسجام يابد. براي مثال داده ها بايد مطابق با مدل داده اي يكپارچه" همگن" شوند. به علاوه، مقادير داده اي سيستم هاي عملياتي بايد پاك شوند تا داده هاي صحيحي در انبار داده ها وارد شوند. نياز به داده هاي تاريخي يكي از موارد مهم در شيوه انبار داده هاست .

داده هاي تاريخي براي تحليل روند كسب و كار ضروري هستند. البته هزينه نگهداري اين گونه داده ها نيز بايد مورد توجه قرار گيرد. به علاوه، داده هاي انبار داده هاي ثابتي هستند، براي مثال دسترسي به DWH از نوع خواندني است. انجام اصلاحات در اين داده ها فقط هنگامي صورت مي گيرد كه اصلاحات داده هاي منبع در انبار انتشار يابند DWH .داده هاي ديگري به نام داده هاي اشتقاق يافته (derived data) دارد. اين داده ها به طور صريح در منابع عملياتي ذخيره نمي شوند، بلكه در حين بعضي از فرآيندها از داده هاي عملياتي، اشتقاق مي يابند. براي مثال داده هاي فروش را مي توان در سطوح مختلف (هفتگي، ماهانه، فصلي) در انبار ذخيره نمود .

سيستم هاي انبار داده ها

سيستم انبار داده ها و همه مولفه هايي است كه براي ساخت، دستيابي و نگهداري DWH به كار مي روند. انبار داده ها بخش مركزي سيستم انبار داده ها را تشكيل مي دهد. گاهي اوقات انبار داده ها حجم عظيمي از اطلاعات را در واحد هاي منطقي كوچكتر به نام Data Mart نگهداري مي كند .مولفه آماده سازي، مسوولیت كسب يا دريافت داده ها را بر عهده دارد. اين مولفه شامل همه برنامه ها و برنامه هاي كاربردي اي است كه مسئول استخراج داده ها از منابع عملياتي هستند. مولفه دستيابي شامل برنامه هاي كاربردي مختلف OLAP) يا برنامه هاي كاربردي داده كاوي) است كه امكان استفاده از اطلاعات ذخيره شده در انبار داده ها را فراهم مي آورند .

مولفه مديريتMetadata ، وظيفه مديريت، تعريف و دستيابي به انواع مختلف Metadata را به عهده دارد. در اصل، Metadata"داده هايي درباره داده ها" يا "داده هايي است كه مفهوم داده ها را توصيف مي كنند". انواع مختلف Metadataدر انبار داده ها وجود دارند. مثلا اطلاعاتي در مورد منابع عملياتي، ساختار داده هاي DWH و كارهايي كه در حين ساخت، نگهداري و دستيابي به DWH انجام مي شوند. نياز به Metadata شناخته شده است. پياده سازي يك DWS منسجم، كار پيچيده و دشواري است و شامل دو فاز مي باشد. در فاز اول كه پيكربندي DWS نام دارد، ديدگاه مفهومي انبار داده ها مطابق با نيازمندي هاي كاربر مشخص مي شود. سپس منابع داده اي دخيل و روش استخراج و بارگذاري در انبار داده ها تعيين مي گردد. سرانجام، درباره پايگاه داده هاي مورد نظر و روش هاي دستيابي داده ها تصميم گيري خواهد شد. پس از بارگذاري اوليه، در فاز عمليات DWS بايد داده هاي انبار داده ها به منظور منظم refresh شوند.

طراحي انبار داده ها روش هاي طراحي انبار داده ها امكان پردازش كارآمد query را بر روي حجم عظيمي از داده ها فراهم مي آورند. نوع ويژه اي از الگوي پايگاه داده ها به نام star براي مدل سازي انبار داده هاي چند بعدي به كار مي رود .در اين حالت، پايگاه داده ها از يك جدول مركزي واقعيت يا fact و جداول چند بعدي تشكيل شده است. جدول واقعيت حاوي tuple هايي است كه بيانگر واقعيت هاي كسب و كار مانند فروش يا عرضه هستند. هر tupleجدول واقعيت به tuple هاي جدول چند بعدي اشاره دارد. هر tuple جدول چند بعدي مواردي نظير محصولات، مشتريان، زمان و فروشنده را نشان مي دهد.

انبار داده هاي مجازي

هدف انبار داده هاي مجازي، پياده سازي سريع انبار داده ها بدون نياز به ذخيره سازي و نگهداري كپي هاي متعدد از داده هاي منبع است. اغلب، انبار داده هاي مجازي به سازمانها كمك مي كند تا به نياز واقعي كاربران نهايي پي ببرند. كاربران نهايي مي خواهند به طور مستقيم به داده هاي منبع بي درنگ با كمك ابزارهاي توانمند شبكه اي دسترسي پيدا كنند. معايب اين روش عباتند از :

-كيفيت و سازگاري داده هاتضمين نمي شود. زيرا فعاليت هاي آماده سازي داده ها صورت نمي گيرند .
-به طور معمول، داده هاي تاريخي وجود ندارند .

-زمان دسترسي كاربرنهايي بسته به وجود يا عدم وجود منابع عملياتي، بار شبكه و پيچيدگي درخواست، غير قابل پيش بيني است.

معماري انباره داده از سه لايه تشكيل شده است:

-در اولين لايه اين معماري، سرويس دهنده انباره داده‌اي است كه يک سيستم پايگاه داده رابطه‌اي مي باشد . اين لايه داده هاي مورد نياز خود را از داده هاي عملياتي و منا بع خارجي و فايلهاي مسطح و غيره براي ايجاد انباره داده استخراج مي كند.

-در لايه مياني يک سرويس دهنده پردازش تحليلي برخط مي‌باشد كه بوسيله آن مي‌توان مكعبهاي چند بعدي ساخت . پردازش تحليلي برخط يک ابزار قدرتمند، سريع و مناسب براي گزارشگيري مي‌باشد.

-در آخرين لايه ما ابزارهاي گزارش گيري و تحليل و داده كاوي را داريم .

براي پياده‌سازي يك انباره داده بايد هريك از اين لايه‌ها به درستي پياده‌سازي شوند.

اخذ داده

اخذ داده از منابع مربوطه (پايگاه داده منبع (انجام مي‌گردد. اين مرحله بخش استخراج اطلاعات (Extract)از سري عمليات ETL است .براي انجام عمليات اخذ داده، بايد منبع اخذ داده، نحوه اخذ داده، فرمت داده‌هاي اخذ شده و مقاطع زماني اخذ داده‌ها، همچنين نحوه دسترسي به اين داده‌ها معلوم و مشخص باشد.

بررسي و پاكسازي داده ها

اين مرحله بخش تغيير شكل (Transform) از عمليات ETL است. بررسي و پاكسازي داده هاي استخراج شده جهت ورود به انباره داده در اين مرحله انجام مي‌گيرد. پس از بررسي جداول موجود، فيلدهاي موجود در جداول و محتويات فيلدهاي مذكور، كليه مشكلات داده‌اي در قالب ليستي ارائه مي‌گردند. سپس عمليات پاكسازي براي آنها انجام مي‌گيرد. اين عمليات غالبا در زمره يكي از موارد زير هستند :

-حذف مقاديرnull

-هم مقدار سازي فيلدهاي مشابه از نظر معنا

-ايجاد فيلدهاي كمي جديد قابل بدست آمدن از روي داده‌هاي جدول و مورد نياز

-يكي كردن داده‌ها از منابع مختلف

-خلاصه سازي سطرهاي هم معني كه ايجاد افزونگي مي‌كنند.

-ايجاد كليد جانشين براي جداول

) Pivoting -تبديل چند ستون به چند سطر يا بالعكس(

-تقسيم يك ستون جدول به چند ستون

طراحي انباره داده موضوعي

انباره داده موضوعي: از آنجا كه كاربران مختلف با نيازهاي متفاوتي وجود دارند كه مي‌توانند از داده‌هاي درون انباره داده استفاده كنند، برآوردن نيازهاي تمام كاربران به وسيله يك سيستم مركزي هميشه امكان پذير نيست. از طرفي يك سيستم مركزي، متمركز بر روي داده و سيستم مي‌باشد و كاربر نهايي ممكن است كه بخواهد كنترل بيشتري روي محيط اطلاعاتي خود داشته باشد .

راه حل اين مشكلات مركز داده اي است، كه به آن انباره داده‌اي سازماني نيز گفته مي‌شود. مركزداده اي، انباره داده خاصي است كه داده‌هاي مورد نياز براي يك بخش از سازمان يا كاربرهاي مرتبط به آن را جمع آوري مي‌كند.

طراحي مراكز داده‌اي مربوطه بنا به صلاحديد فرد خبره در قالب مدل ستاره‌اي يا دانه‌برفي و يا طرح منظومه حقايق صورت مي‌گيرد.

طرح ستاره‌اي: عمومي‌ترين نمونه براي مدلسازي مدل چند بعدي، طرح ستاره است. در اين طرح انباره داده شامل يک جدول بزرگ مركزي به نام جدول حقايق و يک سري جدول كوچكتر به نام جدول بعد يا جدول بعد كه وابسته به جدول حقايق هستند مي‌باشد.

طرح دانه برفي: اين طرح، تغييريافته طرح ستاره‌اي است بطوريكه بعضي از جداول بعد، نرمال شده‌اند. تفاوت اصلي بين طرح ستاره‌اي و طرح دانه برفي اين است كه جدول بعد در طرح دانه‌برفي به فرم نرمال نگهداري مي‌شود تا ميزان افزونگي كاهش پيدا كند. اين كار باعث كاهش ميزان حافظه مورد نياز خواهد شد. البته صرفه‌جويي در فضاي ذخيره‌سازي جدول بعد در مقايسه با حجم جدول حقايق ناچيز است چون تعداد اتصالاتي كه براي پردازش يک گزارش بايد گذارده شود در اين حالت افزايش مي‌يابد و مدت زمان پاسخ‌دادن به گزارش در مدل دانه برفي بيشتر از مدت زمان لازم در مدل ستاره‌اي است بنابراين غالبا طرح دانه برفي در طراحي انباره داده عموميت طرح ستاره‌اي را ندارد مگر آنكه بنا بر صلاحديد فرد خبره بر طرح ستاره‌اي ترجيح داده شود.

طرح منظومه حقايق: هنگامي كه نياز به چندين جدول حقايق وجود دارد كه داراي جداول بعدهاي مشترک هستند، طرحي ايجاد مي‌شود كه به آن طرح كهكشان يا منـظومه حقايق مي‌گويند. يک طرح منظومه حقايق به جداول ابعاد اجازه مي‌دهد كه بين جداول حقايق مشترک باشند.

وارد سازي داده‌هاي پاكسازي شده به انباره داده

با توجه به فرمت داده‌هاي اخذ شده، وارد سازي داده‌هاي پاكسازي شده به انباره داده با اجراي اسكريپت مربوطه - بخش Load از عمليات.ETL

تحليل داده با ابزار پردازش تحليلي برخط و داده‌كاوي

عمليات پردازش تحليلي برخط بر انباره‌داده‌هاي موضوعي اعمال مي‌گردد. بسياري فكر مي كنند كه داده كاوي و OLAP دو چيز مشابه هستند در اين بخش سعي مي كنيم اين مسئله را بررسي كنيم و همانطور كه خواهيم ديد اين دو ابزار هاي كاملا متفاوت مي باشند كه مي توانند همديگر را تكميل كنند.

OLAPجزيي از تكنيك‌هاي تصميم گيري مي باشد. سيستم هاي سنتي گزارش گيري و پايگاه داده اي آنچه را كه در پايگاه داده بود توضيح مي دادند حال آنكه در OLAP هدف بررسي دليل صحت يک فرضيه است].Error! Reference source not found [.

بدين معني كه كاربر فرضيه اي در مورد داده ها و روابط بين آنها ارائه مي كند و سپس به وسيله ابزار OLAPبا انجام چند Query صحت آن فرضيه را بررسي مي كند.

اما اين روش براي هنگامي كه داده ها بسيار حجيم بوده و تعداد پارامترها زياد باشد نميتواند مفيد باشد چون حدس روابط بين داده ها كار سخت و بررسي صحت آن بسيار زمانبر خواهد بود.

تفاوت داده كاوي با OLAP در اين است كه داده كاوي برخلاف OLAP براي بررسي صحت يک الگوي فرضي استفاده نمي شود بلكه خود سعي مي كند اين الگوها را كشف كند.

درنتيجه داده كاوي و OLAP مي توانند همديگر را تكميل كنند و تحليل گر مي تواند به وسيله ابزار OLAP يک سري اطلاعات كسب كند كه در مرحله داده كاوي مي تواند مفيد باشد و همچنين الگوها و روابط كشف شده در مرحله داده كاوي مي تواند درست نباشد كه با اعمال تغييرات در آنها مي توان به وسيله OLAP بيشتر بررسي شوند.

با گسترش شگرف اينترنت و استفاده روزافزون از آن در جهت ارايه و يا كسب اطلاعات، شاهد حجم انبوهي از اسناد و مقالات بر- خط هستيم كه بعنوان يكي از مشخصات بارز زندگي مدرن امروزي، تحت عنوان افزونگي اطلاعاتي مطرح مي‌گردد. در اين ميان دسترسي سريع و صحيح به منابع مهم و مورد علاقه، يكي از دغدغه هاي استفاده كنندگان از اين منبع اطلاعاتي بسيار بزرگ است. آنچه امروزه از اهميت بسيار زيادي برخوردار گرديده ، كمبود يا نبود اطلاعات نيست بلكه كمبود روشهايي در جهت يافت و بهره برداري از اطلاعات در دسترس به نحوي بهينه است. بعنوان مسئله اي آرماني تر به دنبال روشهايي هستيم تا از اطلاعات موجود به كسب دانش پرداخته،

احتمالاً به ارايه مسايل جديدي بپردازد كه قبل از آن مشخص نبوده است.
متن كاوي ، بعنوان روشي در استخراج دانش از متون، يكي از موضوعات مهم در گستره اي از اعمال مديريت اطلاعات است. در اين ميان آنچه از اهميت فوق العاده اي برخوردار است ارايه راه‌كارهايي براي مواجه با اين حجم عظيم اطلاعاتي و استفاده بهينه از اطلاعات در جهت خلق دانش، توليد سينرجي و در نهايت افزايش خرد جمعي است.
در سالهاي اخير اهميت متون به عنوان منابع با پتانسيل اطلاعاتي بسيار بالا به نحو گسترده‌اي مورد توجه قرار گرفته به طوري كه كشف دانش از متون به عنوان يكي از مهمترين فعاليتهاي محققين حوزه هوش مصنوعي و فناوري اطلاعات قرار گرفته است. تحقيقات بسياري صورت گرفته اما محدوده فعاليت بقدري گسترده است كه نيازمند توجه بيشتري مي‌باشد.
امروزه محققان به اين مسئله معترفند كه با وجود انجام تحقيقات بي وقفه در زمينه كاري خود، نمي‌توانند همزمان با پيشرفت دانش، معلومات خود را به روز نگاه دارند. بعنوان مثال بانك اطلاعاتي Medline در حال حاضر حاوي ١٠ ميليون چكيده مقاله است و هر هفته بين هفت تا هشت هزار چكيده مقاله به اين بانك اطلاعاتي افزوده مي‌شود. در اين بين شايد همه مقالات مربوط به يك دانش خاص نباشند، اما تعداد مقالات تخصصي كه در حوزه تحقيق يك دانش خاص قرار مي‌گيرد به اندازه اي است كه يك نفر نمي‌تواند ادعا كند همه آنها را مطالعه كرده است بعلاوه نقش مطالعات عميق و گسترده و استخراج ايده ها و دانش جديد از مطالب مطالعه شده بر كسي پوشيده نيست.
در اين ميان اينترنت بعنوان بزرگترين منبع اطلاعاتي همگاني، تشكيل يافته از صد ها ميليون صفحه اطلاعات است كه به جهت همگاني بودن آن و نبود آينده‌نگري كافي در زمان تشكيل و رشد آن ، متحمل نگاهداري اطلاعات نويسندگان، محققان ، انديشمندان و غيره به همان نحوي كه آنها مي نوشتند گرديد. نبود يك استاندارد همه جانبه و دقيق در تنظيم متون و قرار گيري اين مجموعه عظيم بصورتي غير ساختيافته و يا بعضاً نيمه ساختيافته، جامعه اطلاعاتي را دچار نوعي سردر گمي و مشكل در دستيابي به اطلاعات مورد نياز كرده بطوري‌كه براي يافتن مطالب مورد نظر خود متحمل هزينه هاي زماني بسياري مي‌گردند. محققان به ارايه راه كارهايي براي ساخت يافته كردن اطلاعات نمودند و با ارايه زبانهاي نشانه گذاري استاندارد نظير XML تا حد زيادي جلوي اين از هم پاشيدگي اطلاعاتي را گرفتند اما آنچه همچنان باقي است وجود بسياري از متون غير ساخت‌يافته مي‌باشد؛ در همين راستا ارايه ابزاهايي كه با بررسي متون بتوانند تحليلي روي آنها انجام دهند منجر به شكل گيري زمينه‌اي جديد در هوش مصنوعي و فناوري اطلاعات گرديده كه به يادگيري متن معروف است.
اين حوزه تمام فعاليتهايي كه به نوعي به دنبال كسب دانش از متن هستند را شامل مي‌گردد. آناليز داده هاي متني توسط تكنيكهاي يادگيري ماشين، بازيابي اطلاعات هوشمند، پردازش زبان طبيعي يا روشهاي مرتبط ديگر همگي در زمره مقوله يادگيري متن قرار مي‌گيرند. يكي از روشهايي كه ذكر گرديد، استفاده از تكنيكهاي يادگيري ماشين در زمينه پردازش متن است، مسئله قابل تامل اين است كه اين تكنيكها در ابتدا در مورد داده هاي ساخت‌يافته به كار گرفته شدند و علمي به نام داده كاوي را بوجود آوردند. داده هاي ساخت‌يافته به داده هايي اطلاق مي‌گردد كه بطور كاملاً مستقل از همديگر ولي يكسان از لحاظ ساختاري در يك محل گردآوري شده اند. انواع بانكهاي اطلاعاتي را مي‌توان نمونه هايي از اين دسته اطلاعات نام برد. در اينصورت مسئله داده كاوي عبارت از كسب اطلاعات و دانش از اين مجموعه ساخت يافته. اما در مورد متون كه عمدتاً غير ساخت يافته يا نيمه ساخت يافته هستند ابتدا بايد توسط روشهايي ، آنها را ساختارمند نمود و

سپس از اين روشها براي استخراج اطلاعات و دانش از آنها استفاده كرد. به هر حال استفاده از داده كاوي در مورد متن خود شاخه اي ديگر را در علوم هوش مصنوعي بوجود آورد به نام متن كاوي . از جمله فعاليتهاي بسيار مهم در اين زمينه، طبقه بندي (دسته بندي) متن مي‌باشد.
طبقه بندي متن، يعني انتساب اسناد متني بر اساس محتوي به يك يا چند طبقه از قبل تعيين شده، يكي از مهمترين مسايل در متن كاوي است؛ مرتب كردن بلادرنگ نامه هاي الكترونيكي يا فايلها در سلسله مراتبي از پوشه ها، تشخيص موضوع متن، جستجوي ساختيافته و/ يا پيدا كردن اسنادي كه در راستاي علايق كاربر ميباشد، از جمله كاربردهاي مبحث طبقه بندي (دسته بندي- كلاسه بندي) متن است. در بسياري از موارد ، افراد حرفه اي آموزش ديده، براي طبقه بندي متون جديد به كار گرفته مي‌شوند. اين فرآيند بسيار زمان بر و پر هزينه است و لذا كاربرد خود را محدود مي‌سازد، به همين منظور علاقه روزافزوني به توسعه فناوري هايي در دسته بندي خودكار متن ابراز ميشود.
در هر حال در جوامع اطلاعاتي امروزي آنچه از اهميت روزافزوني برخوردار است، اطلاعات و تبادل آن است و در اين راستا به توسعه فناوري هاي مرتبط پرداخته مي‌شود، اما يك مرحله كاملاً جديد تر و كاملاً مورد توجه جوامع فرا صنعتي، خلق دانش جديد از اطلاعات قبلي است كه اين جوامع آنرا كليد موفقيت خود در آينده دانسته و به سختي در اين زمينه فعاليت مي‌نمايند. بر ما است تا ضمن ارتقاي فناوري اطلاعات در كشور و ايجاد زير ساختهاي لازمه در اسرع وقت، به اينگونه مسائل جدي‌تر كه در زمره Information High Technology قرار مي‌گيرند، بپردازيم.

شماره۱۸–اسفند ۸۴

مقاله

داده كاوی، مفهوم و كاربرد آن در آموزش عالی

احمد سعیدی
دانشجوی دکتری اقتصاد و مدیریت مالی آموزش عالی

مقدمه

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

هم اکنون در هر کشور، سازمان ها، شرکت ها و . . . برای امور بازرگانی، پرسنلی، آموزشی، آماری و . . . پایگاه داده ها ایجاد یا خریداری شده است، به طوری که این پایگاه داده ها برای مدیران، برنامه ریزان، پژوهشگران و . . . جهت تصمیم گیری های راهبردی، تهیه گزارش های مختلف، توصیف وضعیت جاری خود و . . . می تواند مفید باشد. داده کاوی۱ یا استخراج و کشف سریع و دقیق اطلاعات با ارزش و پنهان از این پایگاه داده ها از جمله اموری است که هر کشور، سازمان و شرکتی به منظور توسعه علمی، فنی و اقتصادی خود به آن نیاز دارد.

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

1 - Data mining

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

سابقه داده کاوی

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

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

پژوهش جدی روی موضوع داده کاوی از اوایل دهه ۹۰ شروع شد. پژوهش ها و مطالعه های زیادی در این زمینه صورت گرفته، همچنین سمینارها، دوره های آموزشی و کنفرانس هایی نیز برگزار شده است. نتایج پایه های نظری داده کاوی در تعدادی از مقاله های پژوهشی آورده شده است. مثلاً سال ۱۹۹۱ پیاتتسکی و شاپیرو ۲ « استقلال آماری قاعده ها در داده کاوی» را بررسی نموده اند. سال ۱۹۹۵ هافمن و نش استفاده از داده کاوی و داده انبار۳ توسط بانک های آمریکا را بررسی نموده و بیان کردند که چگونه این سیستم ها برای بانک های آمریکا قدرت رقابت بیشتری ایجاد می کنند. چت فیلد مشکلات ایجاد شده توسط داده کاوی را بررسی نمود و همچنین مقاله ای تحت عنوان « مدل های خطی غیر دقیق داده کاوی و استنباط آماری » ارایه نمود. هندری نیز دیدگاه اقتصاد سنجی روی داده کاوی را تهیه کرد. در این سال انجمن داده کاوی همزمان با اولین کنفرانس بین المللی «کشف دانش و داده کاوی» شروع به کار کرد. این کنفرانس

2 - piatetsky-shapiro
3- Data warehouse

توسعه یافته چهار دوره آموزشی بین المللی در پایگاه های داده درسال ۱۹۸۹ تا ۱۹۹۴ بود. انجمن مذکور، یک سازمان علمی به نام ACM- SIGKDD را ایجاد نمود. سال ۱۹۹۶ ایمیلنسکی۴ و منیلا۵ دیدگاهی از داده کاوی به عنوان «پرس و جو کننده از پایگاه های استنتاجی۶» را پیشنهاد کردند. فایاد، پیاتتسکی – شاپیرو، اودورأسامی پیشرفت های کشف دانش و داده کاوی را عنوان کردند. در سال ۱۹۹۷ منیلا خلاصه ای از مطالعه روی اساس داده کاوی ارایه نمود. باربارا و همکاران نیز دیدگاه کاهش داده ها روی داده کاوی را در گزارش کاهش داده های نیوجرسی ارایه نمودند. همچنین می توان برای کاربرد داده کاوی

در مدیریت مالی می توان، تحلیل داده های مالی و مدل سازی مالی بنینگاه و چاچ کز و هیگینز ۷ را ملاحظه کرد فریدمن نیز مقاله ای در ارتباط با مفهوم آمار و داده کاوی ارایه نمود. سال ۱۹۹۸ هند ۸ مقاله ای تحت عنوان « داده کاوی : آمار یا بیشتر؟ » ارائه نمود. کلینبرگ۹ پائودیمیتریو و راغان ۱۰ دیدگاه اقتصاد سنجی روی داده کاوی و عملکرد داده کاوی به عنوان یک مسئله بهینه را ارایه نمودند. در این سال نیز کنفرانس های ناحیه ای و بین المللی در مورد داده کاوی برگزار شد که از جمله می توان به کنفرانس آسیا و اقیانوسیه درباره کشف دانش و داده کاوی اشاره کرد. سال ۲۰۰۰ هند و همکاران و اسمیت بحث های مقایسه ای بین آمار و داده کاوی را ارایه کردند. سری و استاوا، کولی، رش پاند و تن استفاده از وب در کاوش داده ها و کاربردهای آن را ارایه کردند. سال ۲۰۰۲ کلادیو کانورسانو و همکاران « مدل آمیخته چندگانه جمع پذیر تعمیم یافته » برای داده کاوی را بررسی نمودند. پائلو و گیانلوکاپاسرون، « داده کاوی ساختارهای پیوند برای مدل رفتار مصرف کننده » را ارایه نمودند.

4- Imielnski
5 - Mannila
6 - Inductive databases
7 - Benninga, Czaczkes, Higgins
8 - Hand
9 - Kleinberg
10 - Paodimitriou , Raghavan

مفهوم داده کاوی

عبارت داده کاوی مترادف با یکی از عبارت های استخراج دانش، برداشت اطلاعات، وارسی داده ها و حتی لایروبی کردن داده هاست که در حقیقت کشف دانش در پایگاه داده ها[11] (KDD) را توصیف می کند. بنابراین ایده ای که مبنای داده کاوی است یک فرآیند با اهمیت از شناخت الگوهای بالقوه مفید، تازه و درنهایت قابل درک در داده هاست. واژه کشف دانش در پایگاه داده ها در اوایل دهه ۸۰ در مراجعه به مفهوم کلی، گسترده، سطح بالا و به دنبال جستجوی دانش در اطلاعات شکل گرفته است. داده کاوی کاربرد سطح بالای فنون و ابزار بکار برده شده برای معرفی و تحلیل داده ها ی تصمیم گیرندگان است. اصطلاح داده کاوی را آمار شناسان، تحلیل گران داده ها و انجمن سیستم های اطلاعات مدیریت به کار برده اند در حالی که پژوهشگران یادگیری ماشین و هوش مصنوعی از KDD بیشتر استفاده می کنند. در ادامه چند تعریف از داده کاوی ارائه می شود.

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

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

۳- « داده کاوی فرآیند کشف رابطه ها، الگوها و روندهای جدید معنی داری است که به بررسی حجم وسیعی از اطلاعات ذخیره

11 - Knowledge Discovery of Database

شده در انبارهای داده با فناوری های تشخیص الگو (مانند ریاضی و آمار) می پردازد». (سایت۱۲ http://www.spss.com)

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

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

۱- درک قلمرو

۲- آماده کردن مجموعه داده ها

۳- کشف الگوها (داده کاوی)

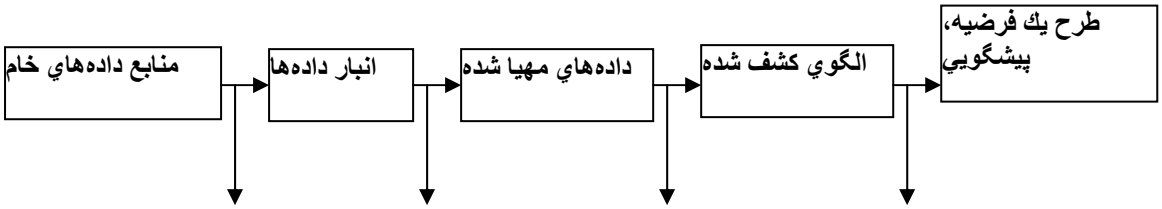
۴- پردازش بعد از کشف الگو

۵- استفاده از نتایج .

فرآیند داده‌کاوی

می‌توان فرآیند داده‌کاوی را طی مراحل زیر به صورت نمودار نشان داد.

۱- این سایت یکی از معتبرترین سایت ها در زمینه آمار و داده کاوی است.



در فرآیند بالا، داده‌های خام از منابع مختلفی جمع‌آوری می‌شوند و از طریق استخراج، ترجمه و فرآیندهای بازخوانی به انبار داده‌ها وارد می‌شوند. در بخش مهیاسازی داده‌ها، داده‌ها از انبار خارج شده و به صورت یک فرمت مناسب برای داده‌کاوی درمی‌آیند. در بخش کشف الگو با روش‌های داده‌کاوی برای پاسخ به سؤال‌های خاصی که به ذهن می‌رسند، الگوریتم‌هایی را استخراج می‌کنند و از این الگوریتم‌ها برای ساخت الگو استفاده می‌شود. در بخش تجزیه و تحلیل الگو، الگوها به یک دانش مفید و قابل استفاده تبدیل می‌شوند و پس از بهبود آن‌ها، الگوهایی که کارا محسوب می‌شوند در یک سیستم اجرایی به کار گرفته خواهند شد.

نرم‌افزارهای داده‌کاوی

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

اکتشاف، فرآیند جستجو در داده‌هاست تا الگوهای مخفی موجود در داده‌ها را بدون هیچ ایدهٔ از پیش تعیین شده‌ای مشخص نماید. در نرم‌افزارهای داده‌کاوی مبتنی بر

مدل‌های پیشگویی، الگوهایی که از یک بانک داده کشف می‌شوند، برای پیش‌بینی آینده به کار می‌روند. مدل‌های پیش‌بینی به کاربر اجازه می‌دهند تا داده‌های نامشخص را به کار ببرد و این مقادیر نامشخص توسط نرم‌افزار کشف شود.

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

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

داده‌کاوی و مدیریت دانش

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

از منظر مدیریت دانش، هدف داده‌کاوی، کشف دانش سازمانی پنهان در اطلاعات خام است. اینگونه نیست که هر بینش حاصل از داده‌کاوی دانش می‌سازد، بلکه در عوض بسیاری از نتایج به دست آمده، اطلاعات مدیریت، یا هوش سازمانی است. مثلاً در سازمان‌های تجاری، دانش با ارزش

مورد مشتری، محصول و بازار را می‌توان از طریق داده‌کاوی به دست آورد. داده‌کاوی ابزار مفیدی برای مدیران دانش است که کشف را با تحلیل تلفیق می‌کنند. تلفیقی که اغلب منجر به ایجاد دانش می‌شود.

کاربرد داده‌کاوی در آموزش عالی

با توجه به اینکه آموزش عالی همواره با داده‌ها و اطلاعات بسیار زیادی در مورد دانشگاه‌ها، دانشجویان، اعضای هیئت علمی، پرسنل، منابع مادی و... روبروست و در اکثر

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

نتیجه گیری

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

منابع:

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

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

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

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

5) Hand. D.J (1998): "Review of Data mining", The American statistician, 52, 112-118.