تکنیک داده کاوی و کاربرد آن در مطالعات اجتماعی

گلمراد مرادی (عضو هیأت علمی دانشگاه آزاد واحد اسلام‌آبادغرب، نویسنده مسؤول)
moradi.pop@gmail.com
وحید قاسمی (دانشیار گروه علوم اجتماعی دانشگاه اصفهان)

چکیده

این مقاله بر موضوع داده کاوی متمرکز است. داده کاوی فرآیندی پیچیده جهت شناسایی الگوها و مدل‌های صحیح، جدید و به صورت بالقوه مفید، در حجم وسیعی از داده ها است، به طریقی که این الگوها و مدل ها برای انسان ها قابل درک باشند. داده کاوی که تا حدودی هم کشف دانش نامیده می‌شود، فرایند تحلیل داده ها از دیدگاه‌های متفاوت و خلاصه کردن آن ها به اطلاعات مناسب می‌باشد. اطلاعاتی که می‌تواند در افزایش سوددهی و تقلیل هزینه ها مفید باشد. این مقاله بر اساس اسناد و مدارک (روش کتابخانه ای) و اطلاعات حاصل از مطالعات انجام شده انجام گرفت. قابل ذکر است که برای بسط بسیاری از مطالب هم از داده‌های موجود استفاده شد ولی در کل روش غالب در این مطالعه توصیفی و کتابخانه ای بود. در مقاله حاضر پس از ارایه خلاصه ای از تفاوت‌های میان روش‌های آماری و داده کاوی، به زمینه‌های استفاده از داده کاوی، استفاده کنندگان، مراحل انجام داده کاوی و روش‌های استفاده از داده کاوی و همچنین محدودیت‌های آن اشاره شده است.

کلیدواژه‌ها: داده کاوی، کشف دانش، اطلاعات آماری، داده.

مقدمه

در سال‌های اخیر در صنایع اطلاعاتی و جوامع مختلف داده کاوی به خاطر دسترسی گسترده به مقدار زیادی از داده ها و نیاز فوری به اطلاعات و شناخت مناسب بسیار مورد توجه قرار گرفته است که این شناخت و اطلاعات به دست آمده برای تحلیل‌های بازار و کاهش کلاهبرداری ها و ضبط و جذب مشتریان مفید بوده است(هان و کمبر[1]، ۲۰۰۰: ۵). در

1- Han & Kamber

مجلهٔ علوم اجتماعی دانشکدهٔ ادبیات و علوم انسانی دانشگاه فردوسی مشهد، بهار و تابستان ۱۳۹۱، صص ۱۵۷-۱۷۸
تاریخ دریافت: ۱۳۸۹/۲/۲۰ تاریخ تصویب: ۱۳۸۹/۸/۲۰

طول دهه گذشته حجم وسیعی از داده ها در پایگاه‌های اطلاعاتی و داده ای جمع آوری شده اند که بیشتر آنها از نرم افزارهای تجاری، درخواست‌های مالی، منابع و مدیریت تجاری و مدیریت روابط مشتری به دست آمده اند. نتیجه این جمع آوری داده‌های حجیم این می‌تواند باشد که سازمان ها و واحدهای تجاری، خدماتی و ... در این راستا اطلاعات غنی ولی شناخت ضعیفی دارند و هدف اصلی داده کاوی این است که از روی داده‌های دم دست نسبت به استخراج الگوها در این مجموعه داده ها اقدام کرده و آنها را به شناخت تبدیل نماید(تانگ و مک کلینان[1]، ۲۰۰۵: ۲). دقت در استخراج داده ها جهت ترسیم وضعیت موجود و وضعیت مطلوب با کمک افراد خبره و کاربران نهایی، موضوع مهمی است منجر به طرح ریزی راهبردی به منظور حصول نتایج مورد نظر می‌باشد. با کمک داده کاوی بر روش پرسشنامه سازمان‌های مشابه و استخراج قواعد و دانش نهفته در آنها می‌توان به معیاری در زمینه ارتباط بین بخش‌های مختلف در زمینه برنامه ریزی استراتژیک دست یافت. به عنوان یک روش برای انجام این کار می‌توان با توجه به معیار وضعیت فعلی پاسخهای مشکوک را استخراج یا با بررسی پاسخهای افراد مربوطه به نتایجی دست یافت(رسولیان و دیگران، ۱۳۸۷: ۷۵). داده کاوی یا کشف دانش در پایگاه داده ها ابزاری فنی و قدرتمند است برای استخراج دانش بالقوه نهفته و اطلاعات پیشین سودمند از مجموعه ای از داده ها به کار می‌رود. این فرایند به صورت خودکار به کشف روابط و الگوهای موجود در داده‌های خام و اجرای نتایج آن می‌پردازد(فرناندز[2]، ۱۹۵۲: ۱۱).

هدف داده کاوی شناخت ارتباطات و الگوهای معتبر، تازه، بالقوه سودمند و قابل فهم از داده‌های موجود می‌باشد. در اذهان عمومی کاوش داده ها به پیدا کردن راه حل اطلاعات سازمان ها و مؤسسات کسب و کار اشاره می‌کند. تعریف یکسانی درباره داده کاوی وجود ندارد. داده کاوی استخراج اطلاعات مفهومی پنهان، ناشناخته و به صورت بالقوه مفید برای مجموعه بزرگ از پایگاه داده ها می‌باشد (لیو و چن[3]، ۲۰۰۹: ۳۵۳۷؛ زوانگ و همکاران[4]، ۲۹۹۰: ٦٦٥ و جرفری[5]، ۲۰۰٤: ۳). که می‌تواند تحلیل‌های ارزشمندی را از میان حجم وسیعی

1- Tang and Maclennan
2- Fernandez
3- Lu & Chen
4- Zhuang et al
5- Jrffrey

از داده ها به دست آورد (جورج[1]، ۱۹۵۲: ۲۲). به عبارت دیگر داده کاوی علم استخراج اطلاعات مفید از پایگاه‌های داده یا مجموعه داده ای می‌باشد (هند و همکاران[2]، ۲۰۰۱). بنابراین داده کاوی استخراج نیمه اتوماتیک الگوها، تغییرات، وابستگی ها، نابهنجاری ها و دیگر ساختارهای معنی دار آماری از پایگاه‌های بزرگ داده ها می‌باشد. بنا به تعریفی دیگر داده کاوی عملیات تجزیه و تحلیلی است که به منظور آشکارسازی نکات نهفته در یک حوزه کاری خاص طراحی شده اند.

توربان و همکارانش داده کاوی را فرایندی می‌دانند که در آن تکنیک‌های آماری، ریاضی، هوش مصنوعی و فراگیری ماشین برای استخراج و شناسائی اطلاعات مناسب و متعاقباً کسب دانش از مجموعه بزرگ از داده ها به کار برده می‌شوند (نای[3]، ۲۰۰۷: ۲۵۹۳). بریسون و همکارانش نیز تعریفی مشابه ارائه دادند. از نظر آنها واکاوی داده ها فرایند استخراج و آشکارسازی الگوهای پنهان از مجموعه ای بزرگ از داده ها است (نای، ۲۰۰۹؛ لژیون[4]، ۲۰۰۱ و احمد[5]، ۲۰۰٤). هم چنین گوبتا به تکنیک‌های اثربخش کاوش و جستجو در داده ها جهت استخراج دانش از آنها داده کاوی می‌گوید (گوپتا[6]، ۲۰۰٦: ۵). هان و کمبر نیز داده کاوی را به استخراج و معدن کاری دانش از مقداری بزرگ از داده ها تعریف کرده اند و آن را شناخت و دانش از داده‌های ناشناخته تعبیر می‌کنند (هان و کمبر[7]، ۲۰۰۰: ۵). براساس تعریف مرکز تحقیقات آمریکا و اداره پاسخگویی سازمان ها داده کاوی به صورت ذیل تعریف می‌شود: داده کاوی مستلزم استفاده از ابزارهای پیشرفته تحلیل برای کشف روابط و الگوهای ارزشمند و ناشناخته در مجموعه ای بزرگ از داده ها است. بر این اساس داده کاوی تنها شامل جمع آوری و مدیریت داده ها نیست، بلکه آن شامل تحلیل و پیش بینی داده ها نیز می‌باشد

1- George
2- Hand et al
3- Ngai
4- Lejeune
5- Ahmed
6- Gupta
7- Han & Kamber

(گزارش داده کاوی[1]، ۲۰۰۶: ۲ و پیتر و همکاران[2]، ۱۹۹۹). این ابزارها شامل مدل‌های آماری[3]، الگوریتم‌های ریاضی[4] و روش‌های یادگیری ماشینی[5] می‌باشند(الگوریتم هایی که خود را به طور ارادی از طریق تجربه مانند شبکه‌های طبیعی[6] و یا درخت تصمیم گیری[7]) بهبود می‌بخشند (پیتر و همکاران، ۱۹۹۹).

با توجه به تعاریف مطرح شده از دیدگاه‌های مختلف، می‌توان دو جزء اساسی را در داده کاوی مشخص نمود. اولی کشف الگوهای پنهان در داده ها می‌باشد و دوم استفاده از این الگوها برای پیش بینی نتایج در آینده است.

در علوم اجتماعی داده ها اغلب حجیم می‌باشند و به آسانی قابل استفاده نیستند، بلکه دانش نهفته ای که در داده ها وجود دارد، قابل استفاده می‌باشد. 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 می‌شود. از آنجائیکه هوش مصنوعی یکی از اصلی ترین عناصر داده کاوی می‌باشد و با توجه به این که به کمک سیستم‌های کامپیوتری و پایگاه‌های داده، روزانه به میزان داده ها افزوده می‌شود، بنابراین استفاده هوشمندانه از دانش بالقوه ای که در این داده نهفته است در دنیای رقابتی امروز برای شرکت ها حیاتی می‌باشد. داده کاوی پیش بینی وضع آینده بازار، گرایش مشتریان و شناخت سلیقه‌های عمومی آنها را برای شرکت ها ممکن می‌سازد.

به طور کلی داده کاوی در تجارت به کار می‌رود و با استفاده از آن شرکت ها می‌توانند داده ها را بطور کامل در مورد الگوها و رفتار مشتریان را برای پیش بینی منافع مورد استفاده، پیش بینی سود و جذب مشتری به کار گیرند(جورج، ۱۹۵۲: ۲۲).

1- Data Mining Report
2- Pieter et al
3- Statistical Models
4- Mathematical Algorithems
5- Machine Learning
6- Natural Networks
7- Decision Trees

هدف اصلی در این مطالعه، توصیفی دقیق از داده کاوی و نقش و کاربرد آن در مطالعات علوم اجتماعی است. در این راستا اهداف فرعی تری مدنظر است که در ذیل بیان می‌شوند:

- زمینه‌های به کارگیری داده کاوی در علوم اجتماعی کدامند؟
- مراحل اساسی انجام داده کاوی کدامند؟
- روش‌های متفاوت در داده کاوی کدامند؟

چرا از داده کاوی استفاده می‌کنیم؟

با توجه به این که در دهه اخیر مؤسسات و سازمان‌های مختلفی توانسته اند مقادیر وسیعی از داده ها را جمع آوری نمایند، داده کاوی می‌تواند به آنها کمک کند که خودشان قادر باشند تا الگوهای پنهان موجود را در داده هایشان استخراج کرده تا بوسیله آن بتوانند استراتژی‌های تجاریشان را گسترش دهند. از طرفی امروزه بیشتر شرکت ها در جهان با نوعی رقابت و چشم و هم چشمی در دنیای تجارت هستند.

در این راستا رمز موفقیت آن ها در این است که برای نگه داشتن مشتریان خود و هم چنین جذب مشتریان جدید از ابزار قوی داده کاوی استفاده کنند. داده کاوی ابزارهایی و فناوریهایی را دربر می‌گیرد که به آنها اجازه می‌دهد تا عواملی را که بر این موضوعات تأثیر می‌گذارند، تحلیل نمایند. در نهایت این که تکنولوژیهای داده کاوی که قبلا تنها در فضای دانشگاهی رایج بود، امروزه این تکنیک دامنه استفاده بسیار وسیعی پیدا کرده و می‌تواند در بسیاری از شرکت ها، صنایع، سازمان ها و مؤسسات خصوصی و دولتی به کار رود(تانگ و مک کلینان، ۲۰۰۵: ٤).

تفاوت بین داده و دانش

داده به هر آنچه که با مشاهدات جهان واقعی در ارتباط است اطلاق می‌شود و ماهیتی در حال تغییر و پویا دارد و کاملا جزئی نگر است در حالیکه دانش به آنچه که کمتر دقیق است و پایداری بیشتری دارد و در ارتباط با تعمیم ها یا تجریدهای داده است، گفته می‌شود. مباحث حاصل از جدول ۱، ارزشمند بودن دانش نسبت به داده را بیان می‌کند که ضرورت روش‌های داده کاوی را برای کشف دانش آشکار می‌سازد. طبق تعریف، به تکنیک‌های اثربخش کاوش و

جستجو در داده ها، جهت استخراج دانش از آنها، داده کاوی می‌گویند(رسولیان و دیگران، ۱۳۸۷: ۷۸).

جدول ۱. ویژگیهای دانش و داده

تفاوت	ویژگیها
داده	• به یک نمونه برمی گردد(یک نمونه از اشیاء افراد و حوادث) • ویژگیها را جداگانه توصیف می‌کند. • با حجم زیاد در دسترس است(آرشیو و بانک اطلاعاتی). • امکان پیش بینی را به ما نمی دهد.
دانش	• به دسته یا مجموعه‌های نمونه برمی گردد. • ساختارها، الگوها و ... را به صورت عمومی توصیف می‌کند. • اغلب کسب آن مشکل است. • قابلیت پیش بینی دارد.

مأخذ: (بیورن[1]، ۲۰۰۲: ۱۸)

پیشینه مفهومی

انتخاب اصطلاح ترکیبی تکنیک داده کاوی[2] ناشی از تشابهات موجود بین جستجوی اطلاعات ارزشمند در یک پایگاه داده بزرگ و کندن صخره ها برای یافتن رگه ای از سنگ معدن باارزش است. هر دوی این کارها بر جستجوی مقدار زیادی از مواد و یا کندوکاو خلاقانه آنها برای تعیین دقیق محل قرار گرفتن چیزهای ارزشمند دلالت دارد(مقدسی، ۱۳۸٤: ۵۱). پژوهش جدی در زمینه داده کاوی از اوایل دهه ۱۹۹۰ شروع شد و پژوهش ها و مطالعات فراوانی در این راستا صورت گرفت و سمینارها، دوره‌های آموزشی و کنفرانس هایی نیز برگزار شده است. استفاده از داده کاوی در سالهای ۲۰۰۰ به بعد در مقایسه با آمار و در سال های ۲۰۰۲ به بعد به عنوان عاملی در رفتار مصرف کنندگان عرضه شد. در گذشته کشف دانش اغلب با استفاده از جمع آوری و ارزیابی داده ها آن هم در مقابل بسیاری پیشفرض‌های

1- Beveren
2- Data Mining

تعریف شده انجام می‌گرفت اما امروزه کشف دانش با استفاده از روش‌های جدیدتری در مقابل رویکردهای قدیمی مانند داده کاوی به عمل می‌آید(راجر و همکاران[1]، ۲۰۰۵: ۳۱۲). با این وجود جای این بحث همچنان در علوم اجتماعی خالی می‌باشد و در مطالعات اجتماعی مطالعات زیادی روی آن انجام نشده است.

زمینه‌های به کارگیری داده کاوی

کاربرد داده کاوی اخیرا بطور وسیعی افزایش یافته است، در حالیکه در ابتدا استفاده کنندگان از داده کاوی اساسا متعلق به صنایع اطلاع بر بودند مانند خدمات مالی و به صورت مستقیم در اختیار بازاریابی بود(جورج، ۱۹۵۲: ۲۲). داده کاوی برای تنوعی از اهداف در دو حوزه خصوصی و عمومی به کار برده می‌شود. صنایعی چون بانکداری، بیمه، پزشکی و به طور عادی از داده کاوی برای کاهش هزینه ها و افزایش فروش استفاده می‌شود. برای مثال در بانکداری و بیمه از داده کاوی برای فاش کردن کلاهبرداری و وضع کردن مالیات استفاده می‌شود(جرفری، ۲۰۰٤: ۳). امروزه عملیات داده کاوی به صورت گسترده توسط تمامی شرکت هایی که مشتریان در کانون توجه آنها قرار دارند، استفاده می‌شود، از جمله فروشگاه ها، شرکت‌های مالی، ارتباطاتی، بازاریابی و غیره. استفاده از داده کاوی به این شرکتها کمک می‌کند 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. هم چنین داده کاوی با استفاده از تقسیم بندی مصرف کنندگان به گروههایی با نیازها و ویژگیهای متفاوت می‌تواند در بازاریابی خدمات ارزنده ای را ارائه دهد(نای و همکاران، ۲۰۰۹: ۲۵۹۳).

یکی از نمونه‌های بارز داده کاوی را می‌توان در فروشگاه‌های زنجیره ای مشاهده نمود، که در آن سعی می‌شود ارتباط محصولات مختلف هنگام خرید مشتریان مشخص گردد. فروشگاه‌های زنجیره ای مشتاقند بدانند که چه محصولاتی با یکدیگر به فروش می‌روند. برای

1- Roger et al

مثال طی یک عملیات پایش داده‌های گسترده در یک فروشگاه زنجیره ای در آمریکای شمالی که بر روی حجم عظیمی از داده‌های فروش صورت گرفت، مشخص گردید که مردانی که برای خرید قنداق بچه به فروشگاه می‌روند معمولا آبجو نیز خریداری می‌کنند. هم چنین مشخص گردید مشتریانی که تلویزیون خریداری می‌کنند، غالبا گلدان کریستالی نیز می‌خرند. نمونه مشابه عملیات داده کاوی را می‌توان در یک شرکت بزرگ تولید و عرضه پوشاک در اروپا مشاهده نمود، به شکلی که نتایج داده کاوی مشخص می‌کرد که افرادی که کراوات‌های ابریشمی خریداری می‌کنند، در همان روز یا روزهای آینده گیره کراوات مشکی رنگ نیز خریداری می‌کنند. به روشنی این مطلب قابل درک است که این نوع استفاده از داده کاوی می‌تواند فروشگاه ها را در برگزاری هوشمندانه فستیوال‌های فروش و نحوه ارائه اجناس به مشتریان یاری رساند.

نمونه دیگر استفاده از داده کاوی در زمینه فروش را می‌توان در یک شرکت بزرگ دوبلاژ و تکثیر و عرضه فیلم‌های سینمایی در آمریکای شمالی مشاهده نمود که در آن عملیات داده کاوی ، روابط مشتریان و هنرپیشه‌های سینمایی و نیز گروه‌های مختلف مشتریان بر اساس سبک فیلم ها(ترسناک، رمانتیک، حادثه ای و...) مشخص گردید. بنابراین آن شرکت به صورت کاملا هوشمندانه می‌توانست مشتریان بالقوه فیلم‌های سینمایی را بر اساس علاقه مشتریان به هنرپیشه‌های مختلف و سبک‌های سینمایی شناسایی کند. به طور کلی داده کاوی در زمینه‌های ذیل به کار می‌روند:

۱- در زمینه‌های تجاری(بازاریابی هدف، تحلیل و مدیریت بازار، تحلیل سبد بازار، پیش بینی قیمت نفت، فهم رفتار مشتری و تحلیل و مدیریت ریسک) با هدف کاهش هزینه پست با موقعیت یابی گروهی از مصرف کنندگان. داده کاوی به طور موثری می‌تواند در مورد الگوها و رفتار مشتریان و کاهش کلاهبرداریها، پیش بینی منابع مورد استفاده و افزایش کسب سود مشتری و کنترل و نابودی مشتریان به کار گرفته شود(فرناندز، ۱۹۵۲: ۱۱). به عنوان مثال کشف الگو در خرده فروشی ها برای شناخت تولیدات به ظاهر ناشناخته که اغلب با هم خریداری شده اند، فهرست بندی کالاهای خریداری شده و هم چنین کالاهای به فروش رسیده در این راستا مستلزم

استفاده از روش‌های داده کاوی است(همان، ۱۲). قابل ذکر است که خرده فروشی ها نیز از داده کاوی برای تولیدات به ظاهر ناشناخته که اغلب با هم خریداری شده اند و ارزیابی اثربخشی سهمیه‌ها و ارتقاء (افزایش فروش تولیدات) استفاده می‌شود(جورج، ۱۹۵۲: ٢٤).

۲- در زمینه شناسایی، مدیریت و کشف تخلف: (شناسایی تخلف هایی چون فریب تلفنی، فریب بیمه اتومبیل، کشف حقه‌های کارت اعتباری، کشف تراکنش‌های مشکوک ملی و پول شویی).

۳- در زمینه متن واکاوی: (خلاصه سازی، یافتن متون مشابه و کلمات کلیدی، پالایش نامه‌های الکترونیکی و گروههای خبری و ...).

٤- در زمینه‌های پزشکی: (کشف ارتباط و علائم بیماری، تحلیل آرایه‌های DNA و ساخت تصاویر پزشکی) داده کاوی می‌تواند پیش بینی کند که کدامیک از مشتریان روش‌های جدید را خواهند خرید. لذا الگوهای رفتاری که مشتریان خطرپذیر را مشخص می‌کند(فرناندز، ۱۹۵۲: ۱۳). شرکت‌های داروسازی می‌توانند سوابق فروش اخیر خود را برای شناسایی دکترهای داروساز و تعیین این که فعالیتهای بازاریابی اثر بزرگی روی نتایج کارشان داشته باشد، افزایش دهند. هم چنین داده کاوی می‌تواند پیش بینی کند که کدامیک از مشتریان روش‌های جدید را خواهند خرید(جورج، ۱۹۵۲: ۲۲).

۵- حمل و نقل: سازمان‌های ایالتی و کشوری حمل و نقل می‌توانند میزان کارایی و مدل‌های شبکه ای بهینه سازی را برای پیش بینی هزینه‌های دوره عمر یک محصول از طریق راههای مختلف مانند شوسه، راه آهن و ... پیش بینی کنند(فرناندز، ۱۹۵۲: ۱۳).

٦- تولیدات صنعتی کارخانه ها: کارخانه ها می‌توانند داده کاوی را برای بهبود فرایند فروششان به خرده فروشی ها از طریق لیست افراد مشتری، محموله ها، فعالیت رقبا و قدرت مانور خودشان به کار ببرند(جورج، ۱۹۵۲: ٢٤).

از دیگر زمینه‌های به کارگیری داده کاوی، استفاده بیمارستانها و کارخانه‌های داروسازی جهت کشف الگوها و مدل‌های ناشناخته تأثیر داروها بر بیماری‌های مختلف و نیز بیماران گروه‌های سنی مختلف را می‌توان نام برد. استفاده از داده کاوی در زمینه‌های مالی و بانکداری به شناخت مشتریان پر

خطر و سودجو بر اساس معیار هایی از جمله سن، درآمد، وضعیت سکونت، تحصیلات، شغل و غیره می‌انجامد.

تفاوت داده کاوی و تحلیل‌های آماری

داده کاوی توسط تجهیزات خاصی صورت می‌پذیرد که عملیات کاوش را بر اساس تجزیه و تحلیل مکرر داده ها انجام می‌دهد. داده کاوی با تحلیل‌های متداول آماری نیز متفاوت است. در تحلیل آماری، آمار شناسان همیشه با یک فرضیه شروع به کار می‌کنند. آنها از داده‌های عددی استفاده می‌کنند و به دنبال پیدا کردن رابطه هایی هستند که به فرضیه آنها مربوط است. آنها می‌توانند داده‌های نابجا و نادرست را در طول تحلیل مشخص کنند و در نهایت نتایج کار خود را تفسیر و برای مدیران بیان کنند. اما در داده کاوی به فرضیه احتیاجی نیست. ابزارهای داده کاوی از انواع مختلف داده، نه تنها عددی می‌توانند استفاده کنند. الگوریتم‌های داده کاوی به طور اتوماتیک روابط را ایجاد می‌کنند. داده کاوی به داده‌های صحیح و درست نیاز دارد و این که نتایج داده کاوی ها نسبتا پیچیده می‌باشد و نیاز به متخصصانی جهت بیان آنها به مدیران دارد. داده کاوی اغلب برای ترکیبی از آمارها(هوش مصنوعی و پژوهش هایی برمبنای پایگاه داده ها) مطرح شده بود، آن هم تا این اواخر بعنوان زمینه ای از علائق آماری شناخته و مطرح نشده بود و بعنوان اصطلاحی ناخوشایند در آمار مطرح شده بود (پیگیبون[1]، ۱۹۹۷: ۸).

به عنوان مثال در مورد شناخت کلاهبرداری‌های شرکت بیمه این دو روش نگاه متفاوتی دارند: در روش تحلیل آماری، یک محقق ممکن است متوجه الگوی رفتاری شود که سبب کلاهبرداری بیمه گردد. بر اساس این فرضیه، محقق به طرح یک سری سوال می‌پردازد تا این موضوع را بررسی کند. اگر نتایج حاصله مناسب نبود، محقق فرضیه را اصلاح می‌کند و یا با انتخاب فرضیه دیگری مجددا شروع می‌کند. این روش نه تنها وقت گیر است بلکه به قدرت تجزیه و تحلیل محقق نیز بستگی دارد. مهمتر از همه این که این روش هیچ وقت الگوهای کلاهبرداری دیگری را که محقق به آنها مظنون نشده و در فرضیه جا نداده، پیدا نمی کند. اما در روش داده کاوی، یک

1- Pregibon

محقق سیستم‌های 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 می‌پردازد.

داده کاوی تمام الگوهای غیرعادی را که از حالت عادی و نرمال انحراف دارند و ممکن است منجر به کلاهبرداری شوند را پیدا می‌کند. نتایج داده کاوی حالت‌های مختلفی را که محقق باید در مراحل بعدی پژوهش کند، نشان می‌دهند. در نهایت مدل‌های به دست آمده می‌توانند مشتریانی را که امکان کلاهبرداری دارند، پیش بینی نمایند.

مثال دوم روش تحلیل سنتی: یک محقق می‌خواهد به مطالعه رفتار خرید یک طبقه مشخص از مشتریها(مثلا معلمان بازنشسته) برای طراحی بازار هدف[1] بپردازد. ابتدا محقق از خصوصیات شناخته شده این طبقه مشتری استفاده کرده و سعی می‌کند آنها را در گروهایی ردیف کند. سپس به بررسی رفتار خرید یکسان در هر یک از این گروها می‌پردازد. او این کار را آنقدر انجام می‌دهد تا به گروه بندی مناسب و مورد رضایتی برسد.

اما ابزارهای داده کاوی به مطالعه بانک داده ها برای مشخص کردن تمام گروهایی که الگوی خرید مشخص دارند می‌پردازد. بعد از کاوش آن داده ها، محقق می‌تواند این نتایج را گزارش دهد و یا برای بررسی مجدد به ابزارهای تحلیلی دیگری دهد.

نرم افزار داده کاوی یکی از مهم ترین ابزارهای تحلیل برای داده ها است و به استفاده کنندگان این امکان می‌دهد که داده ها را از ابعاد و زوایای متفاوتی تحلیل و دسته بندی و روابط مشاهده شده بین آنها خلاصه کنند.

مراحل اصلی انجام داده کاوی

داده کاوی فرایندی تحلیلی است که برای کاوش داده ها (معمولا حجم عظیمی از داده ها – در زمینه‌های کسب وکار و بازار) صورت می‌گیرد و یافته‌ها با بکارگیری الگوهایی، احراز اعتبار می‌شوند.

1- Target Market

هدف اصلی داده کاوی پیش بینی است. داده کاوی را "کشف دانش در داده ها[1]" نیز می‌نامند. کشف دانش داده ها دارای مراحل مختلفی می‌باشد که در اینجا به صورت خلاصه بیان می‌شوند:

۱- شناخت مشکل و تعریف هدف موضوع: یکی از مهم ترین عواملی که باعث قصور داده ها می‌شود، عدم تعریف از هدف براساس مشکلات کوته مدت و بلندمدت شرکت هاست. داده کاوی به صورت اصطلاحی روشن و شاخص هدف تجارت را تعریف می‌کند(فرناندز، ۱۹۵۲: ۳۱)، تا جایی که معلوم می‌کند تا چه اندازه شرکت ها انتظار موفقیت دارند و داده کاوی چگونه به آنها کمک می‌کند.

۲- پاک سازی داده ها[2]: کلید موفقیت داده کاوی، استفاده از داده هایی مناسب است. در این مرحله داده‌های غیرمعتبر از مجموعه داده‌های آموزشی خارج می‌شوند. به عبارتی این مرحله برای برطرف کردن خش اطلاعات و داده‌های نامرتبط به کار می‌رود(تانگ و مک لینان، ۲۰۰۵: ٤). داده‌های دارای اطلاعات ناکامل و ... نمونه هایی از داده هایی هستند که باید پاکسازی در مورد آنها انجام گردد. معمولا این‌مرحله با آماده سازی داده ها صورت می‌گیرد که ممکن است شامل پاک سازی داده ها، تبدیل داده ها و انتخاب زیرمجموعه‌هایی از رکوردها با حجم عظیمی از متغیرها باشد(فرناندز، ۱۹۵۲: ۳۱). سپس با توجه به ماهیت مسأله تحلیلی، این‌مرحله به مدل‌های پیش بینی ساده یا مدل های‌آماری وگرافیکی برای شناسایی متغیرهای مورد نظر و تعیین پیچیدگی مدل‌ها برای استفاده در مرحله بعدی نیاز دارد.

۳- انتخاب داده ها: این مرحله هدف کسب نمونه ای معرف با کیفیت مناسب برای مدلسازی است(زوی و همکاران[3]، ۲۰۰۹: ٦٦٥). داده‌های مرتبط به فرایند واکاوی داده ها از سایر داده ها جدا می‌شود. این مبحث را می‌توان بخشی از فرایند کاهش اطلاعات نیز دانست.

٤- تبدیل داده ها: داده ها به قالبی قابل استفاده برای داده کاوی در می‌آیند. از اعمالی که در این مرحله صورت می‌گیرد، می‌توان به خلاصه سازی و یا محاسبه مقادیر تجمعی اشاره کرد. تبدل داده ها تا حدودی متغیرهای پیوسته را در واحدهای استاندارد شده بیان می‌کند(فرناندز، ۱۹۵۲: ۳٤).

1- Knowledg Discovery in data
2-Data Cleansing
3- Zoe et al

۵- انتخاب مدل: انتخاب مدل و الگوسازی هسته اصلی در داده کاوی است. هدف از مدلسازی در داده کاوی برای کشف روابط معتبر بین عناصر داده ها است. این روابط که تا حدودی به الگوها و قوانین بر می‌گردد، می‌توانند برای پیش بینی رفتار آینده مورد استفاده باشند(گزارش داده کاوی، ۲۰۰٦: ۱۰). تشخیص الگوهای صحیح مورد نظر، از سایر الگو ها در این مرحله انجام می‌شود. صحت الگوها بر اساس یک سری معیارهای جذابیت سنجیده می‌شود(تانگ و مک لینان، ۲۰۰۵: ١٦).

٦- ارزیابی مدل: در نهایت قبل از این که مدل گسترش یابد و اجرا شود، باید ارزیابی شده و اعتبار آن سنجیده شود. در این مرحله با استفاده از محاسبات عددی و پارامترهای مختلف مدل‌های متفاوتی به دست می‌آید. ارزیابی مدل به وسیله مجموعه ای از متغیرهای مستقل است که شرط لازم در داده کاوی برای تصدیق و توسعه مدل می‌باشد(همان، ٤).

۷- بازنمایی دانش: بازنمایی دانش را گزارش دهی نیز می‌نامند و آن یک نهر فرعی برای یافته ها در داده کاوی است. در این بخش به منظور ارائه دانش استخراج شده به کاربر، از یک سری ابزارهای بصری سازی استفاده می‌گردد(زوی و همکاران، ۲۰۰۹: ٦٦٦). معمولا در داده کاوی دو نوع گزارش داده می‌شود: گزارش درباره یافته ها(الگوها) و گزارش هایی درباره پیش بینی یا برآورد داده ها(تانگ و مک لینان، ۲۰۰۵: ٤).

کاهش اطلاعات در داده کاوی

کاهش اطلاعات عبارت است از تولید یک مجموعه کوچک تر از داده‌های اولیه که تحت عملیات داده کاوی نتایج تقریبا یکسانی با نتایج داده کاوی روی اطلاعات اولیه به دست می‌دهد. این عمل را می‌توان از طریق حذف خصیصه‌های غیر مرتبط با نوع عملیات پایش داده‌های مورد نظر انجام داد. حذف خصیصه‌های مرتبط که در اثر اشتباه در ارزیابی میزان ارتباط آنها با عملیات داده کاوی انجام می‌گیرد، می‌تواند منجر به ناکارآمدی فرایند داده کاوی و استخراج قوانین ناقص و در نتیجه بی ارزش شود. عدم حذف خصایص غیرمرتبط می‌تواند زمان انجام عملیات داده کاوی را به طرز قابل ملاحظه ای افزایش دهد. سه روش کلی برای انتخاب خصایص مرتبط با واکاوی داده ها وجود دارد:

۱- انتخاب پیش رونده: در هر مرحله خصیصه ای که بیشترین ارتباط را دارد، برگزیده می‌شود.

۲- انتخاب عقب رونده: در هر مرحله خصیصه ای که کمترین ارتباط را دارد، انتخاب و حذف می‌شود.

۳- روش ترکیبی: ترکیب هر دو روش پیش رونده و پس رونده

سلسله مراتب مفهومی نیز روشی برای کاهش تعداد مقادیر ممکن برای یک خصیصه ارائه می‌دهد، اگر چه داده‌های خروجی کلی تر بوده و فاقد برخی جزئیات هستند، اما این داده ها بسیار ساده تر بوده و در سطح تجریدی بالاتری نسبت به داده‌های اولیه قرار دارند.

اطلاعات مورد نیاز برای عملیات داده کاوی

۱- داده‌های مرتبط با فرایند داده کاوی: بانک اطلاعاتی ممکن است شامل تعداد زیادی از رکوردها باشد که تنها بخش کوچکی از آنها با فرایند داده کاوی مرتبط هستند. مشخص کردن این بخش از اطلاعات باید توسط کاربر انجام گیرد.

۲- نوع دانشی که باید استخراج شود: نوع روتین هایی که باید بر روی داده‌های انتخاب شده اعمال شوند، باید مشخص گردد.

۳- دانش زمینه ای: کاربران می‌توانند، با مشخص کردن دانش زمینه ای فرایند داده کاوی را هدایت نمایند، برای نمونه حدس کاربر در مورد رفتار اطلاعات.

٤- معیارهای ارزیابی دانش استخراج شده: این معیارها ممکن است در زمان اجرای فرایند واکاوی داده ها و یا پس از پایان داده کاوی ، روی دانش استخراج شده اعمال شده و بخش ارزشمند دانش را مشخص نمایند.

۵- نحوه ارائه دانش استخراج شده: نمایش دانش و قوانین استخراج شده در قالب‌های مختلفی نظیر جدول، نمودار، درخت تصمیم گیری و ...

روش‌های مختلف داده کاوی

این روشها بطور کلی به دو دسته زیر تقسیم می‌شوند:

۱- الگوریتم‌های یادگیری با نظارت[1]: در الگوریتمهای یادگیری با نظارت، هدف از داده کاوی مشخص است و می‌دانیم که به دنبال چه نوع دانشی می‌گردیم. روش‌های یادگیری با نظارت شامل دو روش رگرسیونی و طبقه بندی می‌باشند:

الف) مدل‌های طبقه بندی: در مدل‌های طبقه بندی محقق می‌خواهد احتمال عضویت طبقه ای به عنوان تابعی از متغیرهای درونداد مشخص نماید. این مدلها برای پاسخ هایی که طبقه بندی شده اند مناسب هستند.

ب) رگرسیونی: در مدل‌های رگرسیونی محقق می‌خواهد برآورد تقریبی از تابع رگرسیونی داشته باشد. در حالیکه مدل‌های رگرسیونی برای متغیرهای پیوسته و دوتایی مناسب هستند. دو نوع از مدل‌های رگرسیونی عبارتند از رگرسیون خطی و رگرسیون لجستیک. رگرسیون خطی متغیرهای پاسخ پیوسته را از روی تابع متغیرهای پیش بین، پیش بینی می‌کند. اما رگریسون لجستیک را برای متغیرهای وابسته جفتی یا ترتیبی به کار می‌بریم.

ج) مسیر تصمیم گیری و طبقه بندی که برای گسترش مسیر تصمیم گیری با شکافتن مجموعه کوچکی از داده ها به کار می‌رود. هدف از مسیر تصمیم گیری مجموعه ای کوچک از داده‌های همگن با توجه به متغیر هدف است. با این وجود متغیرهای اسمی و طبقه بندی همانند متغیرهای پاسخ در مسیر تصمیم گیری می‌توانند مورد استفاده باشد.

د) تحلیل تابع ممیزی: این نیز روشی طبقه بندی است که برای تعیین متغیرهای پیش بین در تمایز بین دو یا چند گروه رخ می‌دهد. تنها متغیرهای مقوله ای اجازه دارند متغیر وابسته باشند ولی متغیرهای طبقه ای و پیوسته می‌توانند به عنوان متغیر پیش بین باشند.

و) آزمون کی دو: این یک روش طبقه بندی است که برای مطالعه رابطه بین سنجش پاسخ‌های مقوله ای و مجموعه بزرگی از متغیرهای پیش بین احتمالی که ممکن است با هم ارتباط داشته باشند، مورد استفاده قرار می‌گیرند(فرناندز، ۱۹۵۲: ۳٤).

از دیگر روش‌های تحلیل در روش‌های یادگیری با نظارت در داده کاوی می‌توان به مدلسازی شبکه عصبی(که هم برای پیش بینی و هم برای طبقه بندی) ضرورت دارند، استفاده کرد.

1- Prediction Method

۲- الگوریتم‌های یادگیری بدون نظارت[1]: در روش‌های یادگیری بدون نظارت، هدف کاملا تعریف شده نیست. مانند خوشه بندی. این روش ها در زمینه‌های بسیاری تحت دامنه وسیعی از نامها در آمده اند نه توزیع بین متغیرهای پیش بین و پاسخ. مهم ترین روش‌های یادگیری بدون نظارت عبارتند از:

الف) تحلیل اجزاء اصلی: در تحلیل اجزاء اصلی ابعاد داده‌های چندمتغیری با استفاده از تبدیل متغیرهای همبسته به متغیرهای ناهمبسته خطی کاهش می‌یابند.

ب) تحلیل عاملی: در تحلیل عاملی عامل‌های پنهان ناهمبسته کمی مقادیر زیادی از واریانس‌های مشترک و جوابگو برای همبستگی مشاهده شده میان داده‌های چندمتغیری که استخراج شده اند، به عهده می‌گیرند.

ج) تحلیل خوشه‌های گسسته: تحلیل‌های گسسته برای ترکیب موردها به گروهها یا خوشه هایی که هر گروه یا خوشه با هم در مشخصه‌های اصلی متجانس بوده، به کار می‌روند.

د) تحلیل‌های پیوستگی و سبد بازار: تحلیل سبد بازار یکی از رایج ترین و مفیدترین انواع تحلیل ها برای بازاریابی است. هدف تحلیل سبد بازار تعیین آن چه که تولید شده و خریداران نیز با هم خرید می‌کنند، می‌باشد(فرناندز، ۱۹۵۲: ۳۷).

معرفی الگوی استخراج داده ها: داده‌های بدون پارامتر(پارامتر آزاد)

بسیاری از الگوریتم‌های داده کاوی نیازمند تنظیم بسیاری از پارامترهای آزاد هستند. یک الگوریتم بدون پارامتر یا پارامتر آزاد ما را از قرار گرفتن در معرض پیش داوریها در مورد مساله در دست بررسی محفوظ می‌دارد و اجازه می‌دهد که خود داده ها با ما صحبت کنند. این روش داده کاوی به صورت تجسس حقیقی اجازه عمل می‌دهد تا این که ما را مجبور کند در معرض فرضیات داده ها قرار بگیریم. یکی از مشکلات الگوریتم پارامتر پنهان این است که تولید مجدد نتایج آزمایشی منتشر شده و فهم توزیع الگوریتم مشکل می‌باشد(کوق و همکاران[2]، ۲۰۰٤: ۲۲).

1- Description Methods
2- Keogh et al

با وجود تعداد زیاد متغیرها، مشکل می‌توان مفهوم نتایج را درک کرد. بنابراین تحلیل گران از ابزارهایی نظیر کاهش ابعاد یا تغییر شکل داده ها استفاده کرده و سعی می‌کنند آنهایی را که تعدادشان تغییر نمی کند، نمایش دهند. این مراحل در مجموع اکتشاف دانش نامیده می‌شود. اکتشاف دانش متضمن عوامل مختلفی است از جمله:

۱- همبستگی یا کشف قواعد وابستگی[1]: که به آن تحلیل سبد بازار نیز می‌گویند، برای شناسایی مجموعه‌های مشترک از مقوله ها و قوانینی برای هدف فروش به کار می‌رود(تانگ و مک لینان، ۲۰۰۵: ٤). هدف همبستگی برقراری روابط میان مقوله هایی که با همدیگر در یک کارکرد مسلم فرض شده اند. تحلیل سبد خرید و برنامه‌های فروش مقطعی نمونه‌های بارزی برای مدل‌های همبستگی هستند(احمد، ۲۰۰٤). ابزارهای مورد استفاده در مدل‌های تحلیل همبستگی محاسبات علت و معلولی و آماری هستند.

۲- طبقه بندی[2]: یکی از مهم ترین مدل‌های معمول در داده کاوی است. در این روش یک نمونه به یکی از چند دسته از پیش تعریف شده دسته بندی می‌شود. هدف آن ساختن مدلی برای پیش بینی رفتار آینده مصرف کنندگان از طریق طبقه بندی پایگاه‌های داده ای در شماری از طبقات از پیش تعریف شده براساس ضوابط معین است. مانند مشتری روزها به چه رستوران هایی می‌رود و چه سفارشی می‌دهد. مهم ترین ابزارهایی که در این زمینه استفاده می‌شوند عبارتند از: شبکه‌های طبیعی، درخت تصمیم گیری و قانون "*اگر این و آن، پس این*"(نای و همکاران، ۲۰۰۹: ۲۵۹۳ و تانک و مک لینان، ۲۰۰۵: ۲).

۳- رگرسیون[3]: پیش بینی یک مقدار متغیر مبنی بر متغیرهای دیگر. نوعی تکنیک تخمین آماری است که برای هر موضوعی از داده ها، ارزشهایی را پیش بینی می‌کند. موارد استفاده آن مانند منحنی برازش، پیش بینی، مدل سازی روابط علی و تست فرضیات علمی در مورد روابط بین متغیرها می‌باشد. از جمله می‌توان به رگرسیون خطی و رگرسیون لجستیک اشاره نمود.

1- Association Rule Discovery
2- Classification
3- Regression

٤- خوشه بندی[1]: که به آن قطعه قطعه سازی نیز می‌گویند(همان، ٤)، موارد درون یک گروه در آن با هم یکسان ولی با موارد خارج از آن گروه متفاوتند. خوشه بندی شامل فرایند گروهبندی داده ها به طبقات و خوشه هایی که موضوعات درون هرخوشه شباهت بسیار بالایی با همدیگر داشته و با دیگر خوشه ها متفاوت می‌باشند(زوی و همکاران، ۲۰۰۹: ٦٦۳). خوشه ها گروه بندیهای دسته‌های داده ای هستند که بر اساس شباهت برخی از معیارها بوجود می‌آیند. دسته بندی وظیفه قطعه قطعه کردن جمعیت‌های ناهمگون به شماری از گروههای همگن را به عهده دارد(کریر و پاول[2]، ۲۰۰۳). گروه بندی اسناد ناشناخته پیشین از جمله موارد خوشه بندی است(جرفری، ۲۰۰٤: ۳۳). تفاوت این روش با طبقه بندی در این است که در ابتدای محاسبات خوشه ها نامعلوم اند و به عبارتی از پیش تعریف شده نیستند. در این روش داده ها براساس روابط منطقی یا اولویت‌های مصرف کننده تنظیم می‌شوند. معروفترین تکنیک ها در خوشه بندی شبکه‌های عصبی و تحلیل ممیزی هستند(نای و همکاران، ۲۰۰۹: ۲۵۹۳).

۵- تحلیل دنباله[3]: که به آن تحلیل زنجیره ای نیز گفته می‌شود، برای الگوهای موجود در یک سری گسسته به کار می‌روند(تانک و مک لینان، ۲۰۰۵: ٦). ترتیب اتفاقاتی که در طول یک دوره زمانی رخ می‌دهد به عبارتی دیگر الگوهایی که در آن یک واقعه خود وقایعی دیگر را به دنبال دارد مانند تولد بچه و خریدن قنداق(جرفری، ۲۰۰٤: ۳). این روش الگوهای دنباله ای مانند سریهای زمانی را دنبال می‌کند(لیو و چن، ۲۰۰۹: ۳۵۳۷). تحلیل دنباله ها روابط یا الگوهای زمانی را نشان می‌دهد. هدف آن بررسی مدل هایی است که در طول زمان ساخته می‌شوند یا برای استخراج و گزارش تغییر مسیرها در طول زمان به کار می‌رود. از جمله ابزارهای مهم تحلیل دنباله ها می‌توان به تئوری مجموعه ها اشاره نمود(نای و همکاران، ۲۰۰۹: ۲۵۹۳). در این روش داده ها براساس روندها و الگوهای منظم رفتاری استخراج می‌شوند. تفاوت میان تحلیل زنجیره ای و مدل‌های همبستگی در این است که مدل‌های زنجیره ای عناصر را جزء به جزء تحلیل می‌کنند اما مدل‌های همبستگی هر مقوله را در یک چرخه خرید برای هم ارزی و مستقل بودن توضیح می‌دهند(تانک و مک لینان، ۲۰۰۵: ٤).

1- Clustering
2- Carrier & Povel
3 Sequential patterns

٦- پیش بینی: پیش بینی مهم ترین و مشهورترین وظیفه داده کاوی است. کشف الگوهایی که در آن می‌توان به پیش بینی منطقی در خصوص فعالیت‌های آینده مبادرت ورزید(جرفری[1]، ۲۰۰٤: ۳). پیش بینی برآورد مقادیر آینده مبتنی بر الگوهای بسط داده شده است که برای مدل سازی و روابط منطقی مدل در آینده بیان می‌شود. پیش بینی تقاضا مثال بازری از مدل پیش بینی است.

مشکلات سیستم‌های داده کاوی

اگر چه واکاوی داده ها می‌تواند به روابط و الگوهای آشکار کمک کند اما نمی توانند به استفاده کنندگان ارزش و اهمیت این الگوها را نشان دهند. به این دلیل الگوهای کشف شده در این راستا وابسته به چگونگی آنها در مقایسه با شرایط جهان واقعی است. محدودیت دیگر در داده کاوی این است که رابطه هایی را که بین رفتار یا متغیرها شناسایی کرده است، لزوما رابطه علی نیست(جرفری، ۲۰۰٤: ۳). به طور کلی دو مشکل اصلی که اکثر سیستم‌های داده کاوی با آن مواجه هستند، عبارتند از: یکی حجم بالای داده‌های آموزشی و بانکهای اطلاعاتی بسیار بزرگ و دوم وجود عدم قطعیت در اطلاعات. داده های عملیاتی موجود در سیستم‌های اطلاعاتی معمولا دارای عدم قطعیت هستند. عدم قطعیت می‌تواند به اشکال مختلفی در پایگاههای داده ظهور کند. بطور کلی عدم قطعیت در سیستمهای پایگاه داده به دو دسته تقسیم می‌شوند:

۱- اطلاعات ناکامل (مقادیر نامشخص): منظور خصیصه هایی است که مقداری برای آنها ثبت نشده است.

۲- اطلاعات ناسازگار: اطلاعاتی که در اثر اندازه گیری نادرست یا بوجود آمدن نویز در داده ها ایجاد شده باشد و مقادیر ثبت شده با مقادیر واقعی برابر نباشند.

برای رفع مشکلاتی که این سیستم ها در برخورد با داده‌های حجیم دارند، معمولا روش‌های زیر استفاده می‌گردند:

۱- طراحی الگوریتم‌های سریع: کاهش پیچیدگیها، بهینه سازی، موازی سازی

۲- کاهش حجم داده ها: نمونه گیری، گسسته سازی، کاهش ابعاد و ...

1- Jrffrey

۳- به کارگیری یک ارائه رابطه ای: استفاده از قابلیتهای ذخیره و بازیابی اطلاعات در پایگاههای داده

نتیجه‌گیری

روشن است که داده کاوی هنوز در مراحل اولیه تکامل خود به سر می‌برد و لذا موقعیت‌های تحقیقاتی بسیاری را به وجود می‌آورد و هنوز کار زیادی است که باید صورت پذیرد. هم چنین داده کاوی حوزه ای متشکل از رشته‌های مختلف و زمینه تحقیقاتی عینی می‌باشد که به خصوص توجه گروههای تحقیقاتی سیستم‌های اطلاعات را به خود جلب کرده است. به علاوه استفاده از ورودی‌های به دست آمده توسط آمارگیران که با داده ها سروکار دارند، نیاز به متخصصین کامپیوتر و پژوهش در عمل می‌باشد. ارتباط با خبرگان سیستم‌های اطلاعاتی که به دنبال به کار آمدن با حدود انسانی و سیستماتیک موجود هستند، نشانگر این موضوع است. این سه مورد منجر به نتایجی می‌شود که به برخی محدودیت‌های مربوط به کاوش داده ها اشاره دارند. نکته دیگر این که از داده کاوی به جای اکتشاف دانش از بانک اطلاعاتی هم استفاده می‌شود. این در زمانی است که از عبارت کاوش داده ها برای روند عملیاتی در کسب نگرش‌های جدید بهره می‌گیرند. هم چنین هنگامی که از اکتشاف دانش برای ایجاد تکنیک‌های جدید استفاده می‌شود، این عبارت مشترکا به جای هم به کار می‌روند.

منابع

رسولیان، محسن، ابوالقاسم شرایعی و محمدباقر فتحی گوهردانی (۱۳۸۷): "نقش داده کاوی بر قواعد انجمنی در مدیریت استراتژیک"، **فصل نامه بصیرت**، سال پانزدهم، شماره ۳۹، صص. ۷٤-۱۰۰.

مقدسی، علیرضا (۱۳۸٤): "Data Mining چیست؟"، ***ماه نامه کنترل کیفیت***، سال دوم، شماره هفتم، صص. ۵۱-۵۵.

Acting C. P. O .(2006): Report to Congression on the impact of Data Mining Technologies on Privacy and Civil Liberities, *Drpartment of Homeland Security*, Washington, July 6, 2006.

Ahmed, S. R. (2004):Applications of data mining in retail business. *Information Technology: Coding and Computing*, Vol. 2, pp. 455–459.

Beveren, J. V. (2002): "A Model of knowledge acquisition that refocuses knowledge management, *Journal of Knowledge Management*, Vol. 1, pp. 18-22.

Carrier, C. G., and Povel, O. (2003): "Characterising Data Mining Software". *Intelligent Data Analysis*, Vol. 7, pp. 181–192.

Chung, H. M. and Paul, G. (2002): "Data Mining Defination, Concerts and Methods", *Journal of Management Information System*, Vol. 16, No. 1, pp. 11-16.

Doke, R. Robert, B. (2001):Data Mining and Classification on the Information Technology, Sloan Management Review.

Fernandez, G. (1952): Data Mining using SAS Applications, Champan and Hall, A CRc Press Company, New York.

Gupta, G. K. (2006): Introduction to Data Mining with case studies, India, Prentic-Hall.

Han, J. and Kamber, J. (2000): Data Mining: Concepts and Techniques, Dartmouth Poblishing Inc, Multiscience Press.

Hand, H. Mannila, P.(2001): Principles of Data Mining. Cambridge, MIT Press.

Jeffrey W. S.(2004):Data Mining: An Overview, Analyst in Information Science and Technology Policy Resources, Science, and Industry Division, Updated December 16.

Khama, J. (2002): "Knowledg Discovery? Data Mining?", *Computer World Journal*, No. 15, pp. 18-19.

Lejeune, M. A. P. M. (2001): Measuring the impact of data mining on churn management. *Internet Research: Electronic Networking Applications and Policy*, Vol. 11, pp. 375–387.

Lu, C. and Ta-Cheng Chen(2009): "A study of applying data mining approach to the information disclosure for Taiwan's stock market investors", *Expert Systems with Applications*, Vol, 36, pp. 3536–3542.

Keogh, E.; Lonardi, S. and Ratanamahatana, C.A.(2004): Towards Parameter-Free Data Mining, Department of Computer Science and Engineering, University of California, Riverside.

Ngai, E.W. and Chau D.C. (2009): "Application of data mining techniques in customer relationship management:A literature review and classification", *Expert Systems with Applications*, Vol, 36, pp. 2592–2602.

Pieter A. and Dolf, Z.(1999): Introduction to Data Mining and Knowledge Discovery, Third Edition (Potomac, MD: Two Crows Corporation, New York: Addison Wesley.

Poon, S. K.; Davis, J. and Choi, B.(2009): "Augmenting productivity analysis with data mining: An application on IT business value", *Expert Systems with Applications*, Vol, 36, pp. 2213–2224.

Pregibon, D. (1997): "Data Mining". *Statistical Computing and Graphics*, Vol. 7, P. 8.

Roger, H. L.; Chua, E. H. and E-Peng L.(2005): "Linear correlation discovery in databases: a data mining approach", *Data & Knowledge Engineering*, Vol. 53, pp. 311–337.

Tang, Z. M. and Maclennan, J.(2005): Data Mining with SQL Server, Wiley Publishing Inc. Indianapolis, Indiana.

Zoe, Y. Z.; Leonid C.; Frada B. and Ken S. (2009): "Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners", *European Journal of Operational Research*, Vol. 195, pp. 662–675.