تئوری و کلیات مبحث داده کاوی

مولف:

نام نویسنده

۱۳۹۹

سرشناسه	**:**
عنوان و نام پدیدآور	**:**
مشخصات نشر	**:**
مشخصات ظاهری	**:**
شابک	**:**
وضعیت فهرست‌نویسی	**:**
موضوع	**:**
موضوع	**:**
موضوع	**:**
موضوع	**:**
رده‌بندی کنگره	**:**
رده‌بندی دیویی	**:**
شماره کتاب‌شناسی ملی	**:**

نام کتاب:
عنوان و نام پدیدآور:
ناشر و محل نشر :
نوبت و تاریخ انتشار :
امور رایانه‌ای:
طرح جلد :
شمارگان:
لیتوگرافی و چاپ:
شابک دوره:
شابک جلد اوّل:

بهای تک جلدی: ..... تومان

فهرست مطالب

سخنی با خوانندگان

در سال‌های اخیرتوانایی‌های فنی بشر برای تولید و جمع آوری داده‌ها به سرعت افزایش یافته است. از جمله عواملی نظیر استفاده زیاد از بارکد برای تولیدات تجاری، در اختیار داشتن کامپیوتر در کسب و کار، علوم، خدمات دولتی و پیشرفت در وسائل جمع آوری داده‌ها، از اسکن کردن متن‌ها و تصاویر تا سیستم‌های سنجش از دور ماهواره‌ای، همگی در این تغییرات نقش بسزایی دارند.

بطور کلی استفاده همگانی از اینترنت و وب به عنوان یک سیستم اطلاع رسانی جهانی ما را با حجم انبوهی از داده‌ها و اطلاعات مواجه کرده است. این رشد انفجاری در داده‌های ذخیره شده، نیاز مبرم وجود تکنولوژی‌های جدید و ابزارهای خودکاری را ایجاد کرده که به صورت هوشمند به انسان کمک کند تا این حجم زیاد داده را به اطلاعات و دانش تبدیل کند؛ داده کاوی به عنوان یک راه حل برای این مسائل مطرح می‌باشد. داده کاوی فرآیندی است که بصورت خودکار استخراج الگوهایی که دانش را بازنمایی می‌کند که این دانش به صورت ضمنی در پایگاه داده های عظیم، انبار داده و دیگر مخازن بزرگ اطلاعات، ذخیره شده است. داده کاوی بطور همزمان از چندین رشته علمی بهره می‌برد که نظیر؛ تکنولوژی پایگاه داده، هوش مصنوعی، یادگیری ماشین، شبکه‌های عصبی، آمار،

شناسایی الگو، سیستم‌های مبتنی بر دانش، حصول دانش، بازیابی اطلاعات، محاسبات سرعت بالا و بازنمایی بصری داده می‌باشد.

داده کاوی در اواخر دهه ۱۹۸۰ پدیدار گشته و در دهه‌های بعدی گام های بلندی در این شاخه از علم برداشته شده و انتظار می‌رود روز به روز به رشد و پیشرفت خود ادامه دهد. کشف دانش در پایگاه داده فرآیند شناسایی درست، ساده، مفید و نهایتاً الگوها و مدل‌های قابل فهم در داده‌ها می‌باشد، داده کاوی مرحله‌ای از فرآیند کشف دانش می‌باشد و شامل الگوریتم‌های مخصوص داده کاوی است، بطوریکه، تحت محدودیت‌های موثر محاسباتی قابل قبول، الگوها و یا مدل‌ها را در داده کشف می‌کند. داده کاوی به فرآیند استخراج دانش ناشناخته، درست و بالقوه مفید از داده اطلاق می‌شود. اصلی‌ترین دلیلی که باعث شد داده کانون توجهات در صنعت اطلاعات قرار بگیرد، مساله در دسترس بودن حجم وسیعی از داده‌ها و نیاز شدید به اینکه از این داده‌ها اطلاعات و دانش سودمند استخراج کنیم.

محتوای این کتاب به چند فصل تقسیم شده که هر فصل به معرفی زمینه‌های مهم و کاربردی در بحث داده کاوی پرداخته شده است. آشنایی با مطالب هر فصل برای خواننده ضروری است. در فصل اول، ما به معرفی مقدمه ای از تئوری و کلیات داده کاوی اشاره و سپس مفاهیم کلی از داده کاوی که شامل تاریخچه، اهمیت، مفاهیم کلیدی در داده کاوی را ارائه کردیم. آنچه در فصل‌های بعدی این کتاب ارائه شد؛ شامل الگوریتم‌های مهم و کاربردی، محدودیت‌ها و ماهیت مساله

داده کاوی، نرم افزارهای کاربردی و ابزارها و کاربردهای داده کاوی در صنعت‌های مختلف پرداخته شده است. داده کاوی را می‌توان حاصل سیر تکاملی طبیعی تکنولوژی اطلاعات دانست که این سیر تکاملی ناشی از یک سیر تکاملی در صنعت پایگاه داده می‌باشد که نظیر عملیات جمع آوری داده‌ها و ایجاد پایگاه داده، مدیریت داده و تحلیل فهم داده‌ها که امیدواریم مطالب این کتاب برای خوانندگان عزیز مفید واقع شود و این روند تکاملی در این شاخه همچنان در حال گسترش واقع گردد.

نام نویسنده

تابستان ۹۹

۱

فصل اول

مقدمه

مقدمه

با گسترش فناوری اطلاعات و ارتباطات[1] در جهان و ورود سریع آن به زندگی روزمره مردم مسائل و ضرورت های تازه‌ای بوجود آمده است. امروزه انسان توسعه یافته کسی است که به اطلاعات دسترسی داشته باشد و دسترسی به اطلاعات نه یک ضرورت بلکه یک قدرت محسوب می شود.

در این میان شهرها به عنوان مراکز قدرت انسانی و تمدن های بشری بیش از پیش اهمیت یافته اند. به اعتقاد الوین تافلر[2]، مردم کره زمین تا به امروز سه موج اساسی تحول را پشت سرگذاشته اند:

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

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

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

[1] Information and communication technology
[2] Alvin Toffler

اگر در موج دوم سخت افزارها[1] به کمک انسان ها می آمدند، در موج سوم این نرم‌افزارها[2] هستند که به خدمت بشر می شتابند و تفکرات و تصورات آدمی را به شکل کدهای صفر و یک و با کمک امواج ماهواره ای مبادله می کنند.

در موج سوم، انسان هر روز که بیشتر یاد می‌گیرد، بیشتر می‌فهمد که با حقیقت فاصله دارد. موج سوم را موج خردورزی نیز لقب داده‌اند؛ زیرا در این عرصه‌ها، انسان‌ها دیگر فرصت ندارند زیاد با هم صحبت کنند، همه چیز تعریف شده و برای هر تعریف، یک کد درنظر گرفته شده است.

از سوی دیگر در دنیای به شدت رقابتی امروز، اطلاعات بعنوان یکی از فاکتورهای تولیدی مهم پدیدار شده است. در نتیجه تلاش برای استخراج اطلاعات از داده ها توجه بسیاری از افراد دخیل در صنعت اطلاعات و حوزه های وابسته را به خود جلب نموده است.

حجم بالای داده ها دائماً در حال رشد در همه حوزه ها و نیز تنوع آنها به شکل داده متنی، اعداد، گرافیک‌ها، نقشه‌ها، عکس‌ها، تصاویر ماهواره‌ای و عکس‌های گرفته شده با اشعه ایکس نمایانگر پیچیدگی کار تبدیل داده‌ها 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 داده‌های

[1] Hardware
[2] Softwares

موجود و رسیدن به نتایج معنی دار بکار گرفته شده‌اند. بعلاوه، عملکرد مناسب ابرداده که داده‌ای 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 می‌کنند. این پیشرفت‌ها هم در بعد سخت افزاری و هم نرم افزاری حاصل شده‌اند.

ریزپردازنده‌های[1] سریع، ابزارهای ذخیره داده های انبوه پیوسته و غیرپیوسته، 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 نرم‌افزار را می‌توان بعنوان یک حرکت پیشرونده از ایجاد یک بانک اطلاعات ساده تا شبکه‌ها و بانک‌های اطلاعاتی[2] رابطه‌ای و سلسله مراتبی برای پاسخگویی به نیاز روزافزون سازماندهی و بازیابی اطلاعات ملاحظه نمود. بدین منظور در هر دوره، نظام‌های مدیریت بانک اطلاعاتی مناسب سازگار با نرم افزار سیستم عامل و سخت افزار رایج گسترش یافته‌اند.

در این رابطه می‌توان از محصولاتی مانند، شرکت اوراکل[3]، یونیفای[4]، سیباس[5]، سیستم مدیریت پایگاه داده دی‌بی اس[6], و غیره نام برد. داده‌کاوی[7] یکی از

[1] Microprocessors
[2] Databases
[3] Oracle
[4] Unify
[5] Sybase
[6] Dbase-IV
[7] Data mining

پیشرفت‌های اخیر در راستای فنآوری‌های مدیریت داده‌هاست. داده کاوی مجموعه‌ای 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 می‌کند. انگیزه برای گسترش داده کاوی بطور عمده از دنیای تجارت در دهه ۱۹۹۰ پدید آمد. مثلاً 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، داده‌های هواشناسی و دیدن کانال‌های تلویزیونی از دیگر کاربردهای آن است. در حوزه مدیریت کتابخانه کاربرد داده کاوی بعنوان فرایند ماخذ کاوی نامگذاری شده است. این کتاب به تئوری و کلیات مبحث داده کاوی می‌پردازد.

داده‌کاوی (Data Mining) به مفهوم استخراج اطلاعات نهان و یا الگوها و روابط مشخص در حجم زیادی از داده‌ها در یک یا چند بانک اطلاعاتی بزرگ است. بسیاری از مردم داده کاوی را مترادف واژه‌های رایج کشف دانش از داده‌ها[1] (KDD) می‌دانند. داده‌کاوی، پایگاه‌ها و مجموعه حجیم داده‌ها را در پی کشف و استخراج، مورد تحلیل قرار می‌دهد.

در سال ۱۹۶۰ اصطلاح «صید ماهی[2]» را جهت کشف هر گونه ارتباط در حجم بسیار بزرگی از داده‌ها بدون در نظر گرفتن هیچگونه پیش فرضی بکار بردند. بعد از سی سال و با انباشته شدن داده‌ها در پایگاه داده اصطلاح داده کاوی در حدود

[1] Knowledge Discovery Data
[2] Data Fishing Or Data Dredging

سال ۱۹۹۰ رواج بیشتری یافت. اصطلاحات دیگری که در این پایگاه داده می‌توان به آن اشاره کرد نظیر؛ باستان شناسی داده[۱] یا برداشت محصول[۲] یا کشف اطلاعات[۳] یا استخراج دانش[۴] نیز بکار رفته‌اند.

تفاوت داده با اطلاعات

بسیاری از مردم به اشتباه می پندارند که داده با اطلاعات تفاوت ندارد، و عموماً این دو واژه را بجای هم استفاده می کنند. داده[۵] می تواند هر نوع از کاراکتر شامل متن، عدد، کلمه، صدا و تصویر باشد و در صورتی که توسط انسان مشاهده شود، لزوماً معنای خاصی هم در بر نخواهد داشت. داده‌ها عموماً خام، دسته‌بندی و طبقه‌بندی نشده هستند و در صورتی که بخواهیم از آنها به صورت مستقیم استفاده کنیم، عموماً بی فایده خواهد بود.

حال آنکه پس از طبقه‌بندی، دسته‌بندی و ساختاردهی به داده‌ها اطلاعات[۶] به وجود می آید. می توان از داده ها برای تصمیم گیری و یا ایجاد دانش در مورد یک مقوله استفاده کرد. اطلاعات عموماً برای کاربر مفهوم دارد و قابل استفاده است.

---

[1] Data Archaeology
[2] Information Harvesting
[3] Information Discovery
[4] Knowledge Extraction
[5] Data
[6] Information

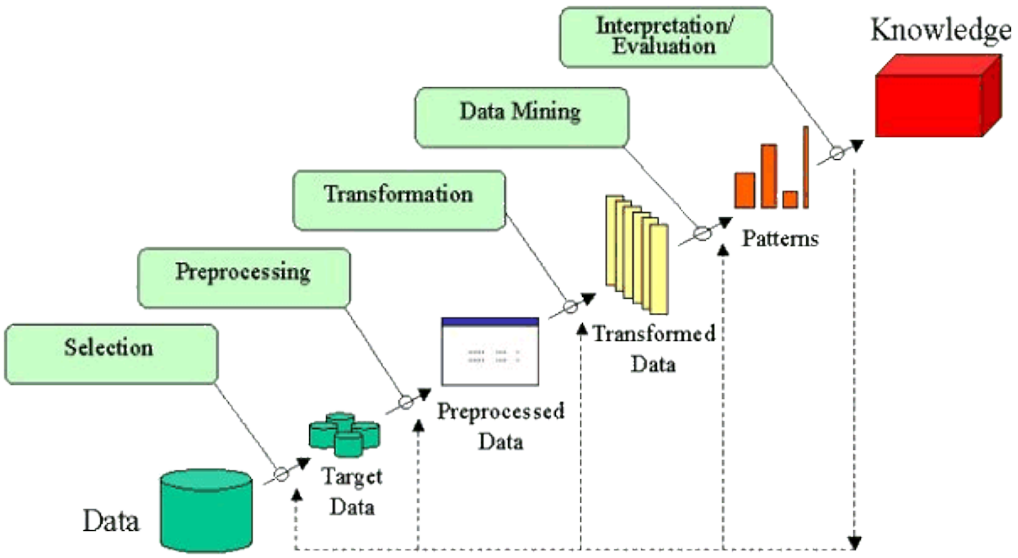
مثال: دانشجویان دانشگاه تهران مرکز واحد مدیریت یک سری داده خام هستند که میانگین نمرات سالیانه آنها نوعی اطلاعات است.

انبوه داده

یک کسب و کار فرضی فعال در زمینه فروش را در نظر بگیرید، این سازمان هر روز صدها مورد فروش را از دهها مشتری ثبت می کند، تمام داده های مربوط به یک خرید از جمله نام و قیمت و دسته کالاها و اطلاعات مربوط به خریدار را ثبت می کند. پس از گذشت مدتی این سازمان انبوهی از داده های بی معنا دارد که نمی تواند از آنها بهره ببرد. این سازمان اگر بخواهد بفهمد کدام مشتریان از چه کالایی بیشتر خوششان آمده راهی ندارد، نمی تواند بفهمد خریداران با هزینه بالا از چه کالایی خرید می کنند و خریداران با سبد قیمتی پایین چه کالایی را می پسندند؛ نمی داند کدام کالا فروش بهتری دارد، چه کالایی در انبار می ماند، چه کالایی حجم سرمایه سازمان را درگیر می کند و در یک کلام: از میان انبوه داده هیچ دانش مفیدی استخراج نمی کند. به همین منظور با استفاده از داده کاوی ارتباط میان اقدامات صورت گرفته و عوامل درونی سازمان مثل قیمت کالاها، تخفیفات، هزینه تبلیغات و دیگر عوامل داخلی را با عوامل بیرونی مثل مشخصات مشتریان (سن، جنسیت، درآمد و محل سکونت)، رقبا و عوامل عمومی بازار (سطح درآمد جامعه، وضعیت رونق و رکود اقتصادی) را می‌توان پیدا کنند. علاوه بر این می توان شاخص هایی مثل رضایت مشتری، درآمد و سود سازمان، مجموع سرمایه درگردش و هزینه های جاری و میزان افزایش و کاهش آنها در طول زمان را استخراج کند.

داده کاوی چیست؟

سازمان ها برای تصمیم گیری و برنامه ریزی به اطلاعات نیاز دارند، بخش مهمی از این اطلاعات از خود سازمان ناشی می شود، از داده‌های قبلی و الگوهای عملکرد سازمان استخراج می شوند، داده‌های خود سازمان نشان‌دهنده رفتار مشتریان و همکاران و بیان کننده موفقیت یا شکست سازمان در یک عمل خاص هستند.



شکل۱- استخراج دانش از میان مجموعه‌ای از داده‌ها با داده کاوی

برای استخراج اطلاعات مفید از میان انبوه حجم داده های ثبت شده باید از فن داده کاوی استفاده کرد. داده کاوی فنی است که از میان پایگاه داده سازمان، به دنبال الگوهای پنهان در میان داده‌ها، ارتباط میان آنها، روند و الگوی آنها می‌گردد. داده کاوی از توابع و الگوریتم های پیشرفته ریاضی استفاده می کند تا ارتباط میان دو دسته از داده و امکان رخ دادن یک نتیجه را در آینده پیش بینی کند.

ویژگی های اصلی داده کاوی

ویژگی[1] یا بُعد[2] در واقع پایه‌ی بسیاری از عملیاتِ داده‌کاوی و یادگیری‌ماشین است. برخی ویژگی های داده کاوی را می‌توان به موارد زیر اشاره کرد؛

- ✓ کشف اتوماتیک الگوها
- ✓ پیش بینی احتمالی نتایج و خروجی‌ها
- ✓ ایجاد اطلاعات اجرایی و مفید
- ✓ تمرکز بر روی داده‌های بزرگ و مجموعه پایگاه‌های داده

[1] Feature
[2] Dimension

داده کاوی چه کاری می تواند انجام دهد؟

شرکت‌ها در یک محیط گسترده از صنایع شامل خرده فروشی، مالی، مراقبت‌های بهداشتی، حمل ونقل و هوا فضا مورد استفاده قرار می‌گیرد، استفاده از داده کاوی با کاربرد الگوی تشخیص تکنولوژی و تکنیک‌های آماری و ریاضی برای تجزیه وتحلیل اطلاعات ذخیره شده از طریق انبار داده‌ها برای استفاده از مزیت‌های داده‌های تاریخی حاصل می‌شود. الگوهای کشف شده و ارتباطات داده‌ها به منظور کمک به تصمیم گیری بهتر کسب وکارها انجام می شود. داده کاوی به شناسایی حقایق مهم، ارتباطات، روندها، الگوها، استثناها و ناهنجاری هایی که ممکن است غیرقابل مشاهده باشند کمک می کند. داده کاوی ممکن است به پیشبرد فروش، توسعه کمپین بازارهای دقیق‌تر، پیش بینی وفاداری مشتریان کمک کند.

بطور مثال شرکت ویدئوی خانگی Blockbuster از داده های سابق مشتریان استفاده میکند و به آنها ویدئوهایی پیشنهاد می دهد تا آنها را تماشا کنند. والمارت (بزرگترین خرده فروش زنجیره ای جهان) برای بهبود عملکرد عرضه کنندگان خود از داده کاوی در مقیاسی وسیع استفاده کرده است. داده‌های ۲۹۰۰ فروشگاه در ۶ کشور برای این کار استفاده شده‌اند و در مجموع ۷٫۵ ترابایت داده مورد بررسی قرار گرفت. ۳۵۰۰ تامین کننده به داده های دسترسی پیدا کردند تا بتوانند الگوهای خرید مشتریان، عملکرد یک کالا و محصول خاص را بررسی کنند و برنامه های خود را بر این پایه و اساس بهبود دهند.

فواید داده کاوی

شناخت مشتریان سودآور: می توانید مشتریانی که بیشترین سود شما از آنها حاصل شده را شناسایی کنید و برای حفظ وفاداری مشتری تلاش کنید.

بهینه‌سازی سبد محصول: شناخت محصولات پر فروش، محصولات سودآور محصولات زیان ده از دیگر فواید داده کاوی است. با این کار می‌توانید در بهتر کردن سبد محصول خود اقدام کنید.

شناخت مشتریان وفادار و قدیمی: می توانید بفهمید مشتریان قدیمی شما چه کسانی هستند و با چه برنامه‌ای خرید می‌کنند، چه کالایی را دوست دارند و چه کالایی باعث وفاداری آنها شده است.

بررسی طول عمر مشتری: با استفاده از داده کاوی می توانید طول عمر مشتری و چرخه آن، میزان سود حاصل عایده از هر مشتری در هر مرحله را بررسی کنید.

شناسایی رفتار مشتری: اگر شما بتوانید رفتار مشتریان خود را بشناسید و آن را با ویژگی های مشتری تطابق دهید می توانید در زمینه بخش بندی و قسمت بندی بازار موفق عمل کنید. اگر امروز بتوانید برای یک محصول خود به صورت مستند بخش بندی بازار انجام دهید در ادامه نیز در این امر موفق خواهید بود.

بررسی عملکرد یک برنامه بازاریابی: اگر می خواهید بدانید یک برنامه بازاریابی و تبلیغاتی که انجام داده‌اید چه اثرات آشکار و پنهانی داشته و برای انتخاب آن در آینده تصمیم بگیرید بی شک داده کاوی بسیار مفید خواهد بود.

کشف الگو[1] **و روند:** با استفاده از داده کاوی و بررسی میزان خرید مشتریان می توانید الگوهای فصلی خرید را استخراج کنید، روند کاهش و یا افزایش آن را تحلیل کنید و در صورت نیاز اقدام اصلاحی انجام دهید.

پیش بینی فروش: با استفاده از اطلاعات گذشته و بهره بردن از الگو و ارتباط میان داده ها می توانید فروش خود را در آینده پیش بینی کنید. روند فصلی فروش را بیابید و برای فروش یک محصول جدید برنامه ریزی کنید.

از جمله نمونه‌های اجرا شده داده‌کاوی در زمینه صنعت می‌توان به موارد زیر اشاره کرد؛

در شرکت‌های خصوصی: شرکت فولادسازی پوهانگ (Pohang) کره جنوبی برای صرفه‌جویی در مصرف انرژی در کوره‌های بلند خود از الگوریتم‌های داده‌کاوی استفاده و در حدود ۱۵٪ از مصرف انرژی خود را کاهش داد. این موضوع چند فایده داشت: ۱/۳ میلیون دلار صرفه‌جویی در هزینه‌های شرکت به ارمغان آورد، قیمت محصولات شرکت را کاهش داد، تعداد مشتریان را افزایش داد و به دنبال آن سود شرکت بیشتر شد.

در صنعت هتل‌داری: یکی از هتل‌های مشهور در لاس‌وگاس آمریکا، برای بالا بردن رضایت مسافران از الگوریتم‌های داده‌کاوی استفاده کرد. به این صورت که با استفاده از اطلاعات جمع‌آوری شده از مسافران به وسیله پرسشنامه، و آنالیز آن داده‌ها توانست عواملی که باعث می‌شد مسافران دوباره به این هتل باز گردند را پیدا کرده و با طبقه‌بندی آنها، مسافران وفادار به هتل را پیدا کنند.

[1] Pattern discovery

در صنعت بانکداری: در یکی از بانک‌های بزرگ کانادا با استفاده از الگوریتم‌های داده‌کاوی، مدلی را برای داده‌ها ارائه داده و به وسیله نتایج آنالیز آن، مسئله مهم تقلب در حساب‌ها و چگونگی و میزان برگشت وام‌های داده شده توسط بانک را حل نمودند و تصمیمی صحیح را برای مشتری‌های جدید بانک گرفتند.

استراتژی و داده کاوی

از آنجائیکه کاربردهای داده کاوی بسیار زیاد است و می تواند در شرکت ها و سازمان ها مختلف متفاوت باشد. می‌توان با استفاده از داده کاوی در حل مشکلات سازمان مورد استفاده قرار گیرد. استراتژیست‌های بزرگ موفق برای مستدل و دقیق بودن استراتژی ها و برنامه ریزی های خود باید از داده کاوی و اطلاعات به دست آمده از آن حداکثر استفاده را ببرند. در واقع برنامه ریزی استراتژیک بدون استفاده از داده کاوی مثل رانندگی با چشمان بسته است! اگر یک استراتژیست می خواهد در زمینه برنامه ریزی کاربردی و اجرایی موفق باشد باید از سلاح داده کاوی استفاده کند.

محدودیت‌های داده کاوی[1]

درحالیکه محصولات داده‌کاوی ابزارهای قدرتمندی می‌باشند، اما در نوع کاربردی کافی نیستند. برای کسب موفقیت، داده کاوی نیازمند تحلیل‌گران حرفه‌ای[1] و

[1] Data mining Constraints

متخصصان ماهری می‌باشد که بتوانند ترکیب خروجی بوجود آمده را تحلیل و تفسیر نمایند. در نتیجه محدودیت‌های داده کاوی مربوط به داده اولیه یا افراد است تا اینکه مربوط به تکنولوژی[۲] باشد.

اگرچه داده کاوی به الگوهای مشخص و روابط آنها کمک می‌کند، اما برای کاربر اهمیت و ارزش این الگوها را بیان نمی‌کند. تصمیماتی از این قبیل بر عهده خود کاربر است. برای نمونه در ارزیابی صحت داده کاوی، برنامه کاربردی در تشخیص مظنونان تروریست طراحی شده که ممکن است این مدل به کمک اطلاعات موجود در مورد تروریست‌های شناخته شده، آزمایش شود. با اینهمه درحالیکه ممکن است اطلاعات شخص بطور معین دوباره تصدیق گردد، که این مورد به این منظور نیست که برنامه مظنونی را که رفتارش به طور خاص از مدل اصلی منحرف شده را تشخیص بدهد.

تشخیص رابطه بین رفتارها و یا متغیرها[۳] یکی دیگر از محدودیت‌های داده کاوی می‌باشد که لزوماً روابط اتفاقی را تشخیص نمی‌دهد. برای مثال برنامه‌های کاربردی ممکن است الگوهای رفتاری را مشخص کند، مثل تمایل به خرید بلیط هواپیما درست قبل از حرکت که این موضوع به مشخصات درآمد، سطح تحصیلی و استفاده از اینترنت بستگی دارد. در حقیقت رفتارهای شخصی شامل شغل (نیاز به سفر در زمانی محدود) وضع خانوادگی (نیاز به مراقبت پزشکی برای مریض)

---

[1] Professional Analysts
[2] Technology
[3] Variables

یا تفریح (سود بردن از تخفیف دقایق پایانی برای دیدن مکان‌های جدید) ممکن است بر روی متغیرهای اضافه تاثیر بگذارد.

ابزارهای داده کاوی[1]

امروزه داده‌کاوی فرآیندی است که برای مدیریت کسب‌وکار بسیار مهم است. استخراج داده‌ها به شدت وابسته به یادگیری ماشین، هوش مصنوعی، سیستم‌های پایگاه داده، تجزیه و تحلیل و الگوریتم‌ها است. به همین منظور یکسری از معروف‌ترین ابزارهای داده‌کاوی که شامل موارد زیر می‌باشد را معرفی و توضیحات تکمیلی درباره آنها در فصل‌های بعدی ارائه داده می‌شود؛

- ابزار Rapid Miner
- ابزار Orange
- ابزار GraphLab Creat
- ابزار R Studio
- ابزار Weka
- ابزار KNIME
- ابزار Apache UIMA
- ابزار CLUTO

[1] Data Mining Tools

- ابزار Anaconda
- ابزار Shogun
- ابزار TraMineR
- ابزار ROSETTA
- ابزار OpenNN

و ...

بهترین ابزارهای داده کاوی به شما اجازه می‌دهد که الگوها را در مجموعه داده‌های بزرگ پیدا کنید. داده‌ها را می‌توان برای توسعه محصولات، خدمات، بهبود محتوای سایت و همچنین برخی از کارهای بازاریابی دیگر استفاده کرد.

الگوریتم‌های داده کاوی

الگوریتم داده کاوی به یکسری روش‌های اکتشافی و محاسباتی گفته می‌شود که هدف آنها ایجاد یک مدل از داده های مورد نظر است. برای ایجاد یک مدل، ابتدا داده‌ها برای یافتن نوعی الگو یا رویکرد توسط الگوریتم تحلیل می‌شود.

سپس الگوریتم با اعمال نتیجه حاصل از این تحلیل بر روی نمونه‌ها، بهینه‌ترین پارامترها را یافته و یک مدل ایجاد می‌کند. سپس این پارامترها بر روی مجموعه داده‌ها اعمال شده و یک الگوی کاربردی به دست می‌آید. به همین منظور، داده‌کاوی شامل الگوریتم‌های متعددی است اما بصورت کلی این الگوریتم‌ها در پنج دسته زیر قرار می‌گیرند؛

۱- الگوریتم‌های طبقه‌بندی[1]

پیش بینی در الگوریتم‌های طبقه‌بندی، براساس یک یا چند متغییر گسسته بر روی سایر ویژگی‌های موجود در مجموعه داده صورت می‌گیرد.

۲- الگوریتم‌های رگرسیون[2]

پیش‌بینی الگوریتم‌های رگرسیون شبیه طبقه‌بندی است با این تفاوت که بر روی متغییرهای پیوسته است. یعنی پیش‌بینی یک یا چند متغییر پیوسته بر روی سایر ویژگی‌های مجموعه داده است.

۳- الگوریتم‌های دسته‌بندی[3]

همانطور که از نام این این الگوریتم‌ها مشخص است، وظیفه‌اش دسته‌بندی است. این الگوریتم‌ها؛ داده‌ها را به گروه و دسته‌های تقسیم می‌کنند که هر دسته دارای ویژگی‌های مشابهی هستند.

۴- الگوریتم‌های وابستگی[4]

کشف روابط و وابستگی میان ویژگی‌های مختلف متغیرها بر عهده این الگوریتم‌ها است. این الگوریتم‌ها به دنبال این هستند که دریابند کدام متغیرها و ویژگی‌ها به هم وابسته هستند و وابستگی آنها به چه شکل است.

[1] Classification algorithms
[2] Regression algorithms
[3] Segmentation algorithms
[4] Association algorithms

۵- الگوریتم‌های تحلیل زنجیره‌ای[1]

این الگوریتم‌ها به دنبال نتایج یکسری رویداد خاص هستند. الگوریتم‌های تحلیل زنجیره‌ای هدفشان کشف توالی و کارهای هست که بصوری سری و توالی انجام می‌شوند تا آنها را تحلیل کنند.

حالا که دسته‌بندی الگوریتم‌های داده‌کاوی ارائه شد؛ به معرفی برخی از برترین الگوریتم‌های داده‌کاوی می پردازیم. در دنیای داده کاوی تعدادی الگوریتم وجود دارد که بسیار پرکاربرد هستند و از قدرت بالایی برخوردار هستد.

این الگوریتم‌ها به شرح زیر می‌باشند:

- الگوریتم K Nearest Neighbor (KNN)
- الگوریتم Classification and Regression Trees (CART)
- الگوریتم Naive Bayes
- الگوریتم Pagerank
- الگوریتم Support Vector Machines
- الگوریتم C 4.5
- الگوریتم K-Means
- الگوریتم Apriori
- الگوریتم Expectation–Maximization

[1] Sequence analysis algorithms

- الگوریتم Support vector machines
- الگوریتم Decision Tree C5
- شبکه Artificial Neural Networks

۲

فصل دوم

مفاهیم داده کاوی