آشنایی با مفاهیم و روش‌های داده‌کاوی

آشنایی با مفاهیم و روش‌های داده‌کاوی

عباس مرادی
جواد حسین زاده
اشکان شباک
کاوه کیانی
محمد شیری

پژوهشکده‌ی آمار

بسم الله الرحمن الرحیم

پیش‌گفتار

در این کتاب روش‌های داده‌کاوی تشریح می‌شود. داده‌کاوی به معنای کشف دانش و استخراج از مقادیر زیادی از داده‌های خام است. از آنجا که تنها ابزار مفید برای استفاده از تحلیل وقایع گذشته در پایگاه داده‌ها به منظور پیش‌بینی در آینده، داده‌کاوی خواهد بود. از این رو برای کشف و استخراج اطلاعات از آمار مد نظر برنامه‌ریزان و مدیران هر سازمانی نیازمند داده‌کاوی خواهیم بود. از سوی دیگر عصر حاضر، عصر روش‌های نوین و متنوع تولید، ذخیره و گرداوری داده‌ها و اطلاعات آماری است. سیاست‌گذاران با حجم انبوهی از اطلاعات آماری متنوع که در کسری از زمان بیشتر و بیشتر تولید می‌شوند در سازمان‌های مهم و کلیدی کشور مواجه هستند. داده‌کاوی اسلوبی برای یافتن دانش و اطلاعات بهینه از میان این انبوه از ریزداده‌ها است. دسترسی به الگوهای مخفی داده‌ها و اطلاعات آماری انباشت شده توسط سازمان‌ها حاکی از آن است که داده‌کاوی بسیار ارزشمند و گران‌بهاست. کتاب حاضر حاوی مفاهیم و روش‌های اساسی داده‌کاوی است و در ۵ فصل تنظیم شده است. در انتهای هر فصل پس از بیان خلاصه به توضیح منابع و مراجع برای مطالعه‌ی بیشتر نیز پرداخته خواهد شد. فصل اول به مقدمه‌ای در مورد

مفاهیم داده‌کاوی می‌باشد. این فصل مشتمل بر چشم‌انداز و ماموریت کتاب برای آشنایی با سایر مطالب کتاب خواهد بود. فصل دوم و سوم به مفاهیم داده‌کاوی و آماده‌سازی (پاک‌سازی و تمیزسازی) داده‌ها اختصاص دارد. در فصل چهارم به پردازش‌های برخط، انبار داده‌ها و پایگاه داده پرداخته‌خواهد شد. در فصل پنجم روش‌های داده‌کاوی شامل درخت تصمیم، خوشه‌بندی، رده‌بندی، قواعد پیوند و سایر الگوریتم‌های داده‌کاوی مورد مطالعه قرار می‌گیرد. فصل آخر به برخی از موضوعات مرتبط با داده‌کاوی پرداخته خواهد شد.

فهرست مطالب

فهرست جداول

فهرست شکل

فصل ۱

مقدمه

۱-۱ مقدمه

داده‌کاوی به معنای کشف دانش و استخراج از مقادیر زیادی از داده‌های خام است. در واقع داده‌کاوی مفیدترین ابزار برای استفاده از تحلیل وقایع گذشته در پایگاه‌های داده‌ای خواهد بود. از این‌رو برای کشف و استخراج اطلاعات از داده‌ها و اطلاعات آماری یک سازمان نیازمند استفاده از داده‌کاوی خواهیم بود. با این اطلاعات کشف شده تصمیم‌گیری‌ها و برنامه‌ریزی‌ها بسیار دقیق‌تر و کاربردی تر خواهد شد. سیاست‌گذاران نیازمند روش‌های داده‌کاوی برای برنامه‌ریزی و سیاست‌گذاری‌های کلان خواهند بود. داده‌کاوی بین محققان علوم آمار، کامپیوتر و ریاضی در جهت تحلیل داده‌ها و پایگاه‌های داده‌ای که شامل مه‌داده‌ها و یا داده‌های با ابعاد بالا هستند، به عنوان یکی از ابزارهای پرتوان شناخته شده است. بر همین اساس مرکز آمار ایران به عنوان یگانه مرجع آمار رسمی کشور روز به روز به اهمیت داده‌کاوی پی برده و این مهم را در اولویت‌های برنامه‌ای خود قرار داده است. پرواضح است که ایجاد یک نظام کارآمد و مؤثر در تولید و عرضه

آمار از الزامات اولیه و ضروری در برنامه‌ریزی است و زیربنای برنامه‌ریزی مناسب، اطلاعات جامع، منسجم و به‌روز است. داده‌کاوی به معنای کشف دانش و استخراج از مقادیر زیادی از داده‌های خام است. از این رو داده‌کاوی مفیدترین ابزار برای استفاده از تحلیل وقایع گذشته در پایگاه‌های داده‌ای مراکز آمار ایران خواهد بود. بدین معنی برای کشف و استخراج اطلاعات از آمار رسمی کشور که آمار مد نظر، سیاست‌گذاران، برنامه‌ریزان و مدیران کشور است نیازمند استفاده از داده‌کاوی خواهیم بود. داده‌کاوی دانشی بین‌رشته‌ای است و ترکیبی از علومی مانند آمار، ریاضیات، کامپیوتر، علوم اطلاعات، نظریه پایگاه داده و یادگیری ماشین می‌باشد. از آنجا که داده‌کاوی یک علم بین‌رشته‌ای است، فلسفه پیدایش آن در درازمدت شکل گرفته است. به طوریکه ابتدا در قرن هجدهم قضیه بیز توسط توماس بیز چاپ و منتشر گردید که بعدها احتمال شرطی نام گرفت. این قضیه از آن جهت مفید است که می‌توان از طریق آن احتمال یک پیشامد را با مشروط کردن نسبت به وقوع و یا عدم وقوع یک پیشامد دیگر محاسبه کرد. این دیدگاه پایه و اساس داده‌کاوی و احتمالات است. قرن بعد [قرن نوزدهم]، قرن پیدایش الگوریتم رگرسیون توسط گاوس بود. ایشان از رگرسیون برای تعیین چرخش اجرام به دور خورشید استفاده کردند. هدف از تحلیل‌های رگرسیونی، یافتن تخمین و روابط مناسب بین متغیرها است. رگرسیون شامل تکنیک‌های زیادی برای مدل سازی و تحلیل متغیرهای خاص و منحصر به‌فرد است. مادامی که هدف یافتن روابط بین متغیر وابسته و یک یا چند متغیر مستقل باشد، می‌توان از این الگوریتم به‌درستی استفاده کرد. تحلیل رگرسیون به صورت گسترده برای پیش‌بینی استفاده شده و یکی از ابزارهای کلیدی در داده‌کاوی به حساب می آید. به کمک رگرسیون پیش‌بینی‌هایی از داده‌های خام برای امور آتی صورت می‌گیرد. در قرن بعد [قرن بیستم]، الگوریتم شبکه‌های عصبی توسط دانشمندان کشف و منتشر شد. این ایده برگرفته از شبکه‌های عصبی مغز بود. درسال ۱۹۴۳ دو پژوهش‌گر به نام‌های وارن مکلوچ و پیتس مقاله‌ای در مورد شبکه‌های عصبی نوشتند [۱۱]. این مقاله اولین و جرقه در زمینه‌ی طراحی شبکه‌های عصبی مصنوعی، برای سایر دانشمندان بود. پس از این مقاله انواع مختلفی از شبکه‌های عصبی مصنوعی توسط اندیشمندان تهیه و تنظیم شد، که همگی برگرفته از ایده‌های این دو دانشمند شهیر بود. تمامی این شبکه‌های

مصنوعی متفاوت و گوناگون دارای یک وجه مشترک بودند که همگی از مغز انسان اقتباس لازم را گرفته‌بودند و نودها و گره‌ها را بر اساس نورون‌های مغز می‌ساختند و یال‌ها و ارتباطات بین گره‌ها را براساس آکسون‌های مغز می‌ساختند. حدود بیست سال بعد یعنی در سال ۱۹۶۶ میلادی شرکت جدیدی به نام علوم تصمیم گیری توسط لورنس فوگل تاسیس گردید. این شرکت در واقع اولین شرکتی بود که به طور ویژه از محاسبات تکاملی برای حل مسائل واقعی استفاده می کرد [۷]. در سال ۱۹۷۰ سیستم‌های مدیریت پایگاه داده پیشرفته اختراع گردید و امکان ذخیره سازی داده‌ها برای همگان فراهم گشت. پنج سال بعد یعنی سال ۱۹۷۵ میلادی، جان هنری هلند کتابی به نام الگوریتم ژنتیک نوشت این کتاب تکنیک جستجویی در علم رایانه برای یافتن راه‌حل تقریبی برای بهینه‌سازی و مسائل جستجو است [۹]. در واقع الگوریتم‌های ژنتیک از اصول انتخاب طبیعی داروین برای یافتن فرمول بهینه جهت پیش‌بینی یا تطبیق الگو استفاده می‌کنند. الگوریتم‌های ژنتیک اغلب گزینه خوبی برای تکنیک‌های پیش‌بینی بر مبنای رگرسیون هستند. در هوش مصنوعی الگوریتم ژنتیک (یا GA) یک تکنیک برنامه‌نویسی است که از تکامل ژنتیکی به عنوان یک الگوی حل مسئله استفاده می‌کند. اما در نهایت پنج سال بعد یعنی سال ۱۹۸۰ میلادی، برای اولین بار اصطلاح «داده‌کاوی» مطرح شد. در این دوره کارشناسان می‌توانستند به روابط معنادار پی ببرند. سپس، در سال ۱۹۸۹ اصطلاح «کشف دانش در پایگاه داده » مطرح شد. در همین زمان اولین کارگاه آموزشی با نام KDD برای استخراج دانش در پایگاه‌های داده گوناگون، شروع به فعالیت نمود. در سال ۲۰۰۱ میلادی، علم داده به عنوان یک رشته مستقل معرفی شد. تا اینکه در قرن حاضر داده‌کاوی به معنای واقعی خود نمایان شد. امروزه داده‌کاوی در اقتصاد، مهندسی، سیاست، پزشکی و... کاربرد فروانی دارد. داده‌کاوی تراکنش‌های مالی، سود سهام، شبکه‌های عصبی، امنیت ملی، الگوریتم‌های ژنتیک نمونه‌هایی از این کاربردهاست. در سال ۲۰۱۴ میلادی آقای حسنی و همکاران به بررسی اهمیت و ارزش کاربردهای داده‌کاوی در آمار رسمی پرداختند [۸].

۱-۲ تعاریف و مفاهیم داده‌کاوی

به‌منظور برداشت یکسان از واژه‌ها و اصطلاحات به‌کار رفته در این کتاب، تعاریف و مفاهیم برخی واژه‌های کلیدی در زیر ارایه می‌شود.

داده[1]: صفت یا اطلاعی است، اغلب عددی، که از طریق مشاهده گردآوری می‌شود. داده‌ها به دو صورت داده‌های خرد و کلان تقسیم‌بندی می‌شوند. در حقیقت، داده‌ها حقایق و واقعیت‌های خامی هستند که اطلاعات از آن‌ها ساخته می‌شوند. این اجزاء در پایگاه‌های داده، ذخیره و مدیریت می‌شوند.

اطلاعات[2]: اضافه کردن زمینه و تفسیر به داده‌ها و ارتباط آن‌ها به یکدیگر، موجب شکل‌گیری اطلاعات می‌شود. اطلاعات در حقیقت داده‌های خلاصه شده را در برمی‌گیرد که گروه‌بندی، ذخیره، پالایش، سازماندهی و تحلیل و تفسیر شده‌اند.

اطلاعات آماری[3]: منظور از اطلاعات آماری، نوع خاصی از اطلاعات هستند که با انجام عملیات‌های ریاضی بر روی داده‌ها و یا دسته‌بندی و طبقه‌بندی آن‌ها بر اساس یک روش تعریف شده به دست آمده‌اند. از این مفهوم برای توصیف بهتر تعریف آمار از بعد عددی آن استفاده می‌شود.

فراداده‌ها[4]: اطلاعاتی درباره‌ی داده‌ها و فرایندهای تولید و نحوه استفاده از داده‌ها است.

پایگاه داده[5]: مجموعه‌ای ساختاریافته از داده‌هاست که در یک سیستم رایانه‌ای نگهداری می‌شود و از راه‌های مختلفی در دسترس قرار می‌گیرد.

دانش[6]: دانش مجموعه‌ای از شناخت‌ها و مهارت های لازم برای حل یک مسأله است لذا اگر اطلاعاتی که در دست است بتواند مشکلی را حل کند می‌توان گفت دانش وجود دارد. ضمن آنکه دانش باید امکان تبدیل به دستورالعمل اجرایی و عملی شدن را داشته باشد.

مدیریت دانش[7]: به معنای در دسترس قرار دادن نظام‌مند اطلاعات و اندوخته‌های علمی است، به گونه‌ای که به هنگام نیاز در اختیار افرادی که نیازمند آن‌ها هستند، قرار گیرند تا

---

[1] Data
[2] Information
[3] Statistical Information
[4] Metadata
[5] Database
[6] Knowledge

آن‌ها بتوانند کار روزمره خود را با بازدهی بیشتر و مؤثرتر انجام دهند. مدیریت دانش شامل یک سری استراتژی و راهکار برای شناسایی، ایجاد، نمایندگی، پخش و تطبیق بینش‌ها و تجارب در سازمان می‌باشد.

داده‌کاوی[۸]: روش‌های کشف الگوها در مجموعه داده‌های بزرگ داده‌کاوی نام دارد. داده‌کاوی دانش و مهارتی است که بین علوم مختلفی همچون یادگیری ماشین، آمار و کامپیوتر قرار دارد که هدف کلی آن استخراج اطلاعات با روش های هوشمند از مجموعه‌ای از داده‌هاست به نحوی که اطلاعات کشف شده به وضوح ملموس‌تر از گذشته باشد.

آمار رسمی[۹]: به اطلاعات عددی گفته می‌شود که توسط دولت یا مراجع صلاحیت دار که در قوانین و مقررات مشخص هستند، تولید و منتشر می‌شود و اطلاعاتی را در مورد وضعیت عمومی کشور برای امور مدیریتی (برنامه ریزی، سیاست گذاری، و تصمیم گیری) به دست می‌دهد. در واقع آمار رسمی به اطلاعات عمومی مربوط می شود که به نفع جامعه و با بودجه دولتی تولید می شود. آمار رسمی برای همگان قابل دسترس خواهد بود.

گرداوری داده[۱۰]: تمام فرایندهایی را شامل می‌شود که مرتبط با بدست آوردن داده از جامعه آمارگیری است.

به‌موقع بودن[۱۱]: به فاصله زمانی بین تاریخ پیش‌بینی شده (اسمی) و تاریخ واقعی انتشار نتایج گفته می‌شود.

بهنگام بودن[۱۲]: به فاصله زمانی بین پایان دوره مرجع و تاریخ واقعی انتشار نتایج گفته می‌شود.

لایه انتشار[۱۳]: برای تولید انتشارات و محصولات آماری به صورت بر خط و برون خط (Online/Offline) و تحلیل و گزارش‌گیری است.

لایه ذخیره‌سازی[۱۴]: برای ذخیره‌سازی داده‌های و فرادادهای معتبر و مرتبط است.

---

[7] Knowledge Managemen
[8] Data Mining
[9] Official Statistics
[10] Data Collection
[11] Punctuality
[12] Timeliness
[13] Dissemination Layer

لایه تولید[۱۵]: برای جمع‌اوری، اعتبارسنجی، پردازش و مدیریت داده و فرادهها است.
انبارداده[۱۶]: مجموعه‌ای از داده‌های موضوع‌گرا، یکپارچه و با تغییرات زمانی است که تصمیمات مدیریتی را حمایت می‌کند.

در واقع تعاریف مختلفی برای انبارداده وجود دارد. دو تعریف مهم و شناخته شده برای انبارداده توسط کیمبال و اینمون ارائه شده است که به آن‌ها در ادامه اشاره می‌شود.

رالف کیمبال -کارشناس زبده‌ی انبارداده -تعریف خلاصه‌ای از یک انبارداده ارایه داده است. در این تعریف وی بر روی وظیفه‌ی یک انبارداده متمرکز شده اما بیل اینمون به چگونگی ایجاد یک انبارداده نیز پرداخته است. از نظر اینمون، انبارداده، نسخه‌ای از داده‌های تراکنشی است که به‌طور ویژه برای انجام پرس و جو و تحلیل ساخت‌یافته شده است.

یکی از متداول‌ترین تعاریف انبارداده توسط بیل اینمون - دانشمند آمریکایی رایانه- ارایه شده است که بسیاری از صاحب‌نظران وی را به‌عنوان پدر این دانش به رسمیت می‌شناسند:

یک انبارداده، مجموعه‌ای از داده‌ها است که به‌صورتی «موضوع‌گرا»، «یکپارچه»، «متغیر با زمان» و «از دست نرفتنی» گرد آمده‌اند و در پشتیبانی فرایند تصمیم‌سازی‌های مدیریتی جای می‌گیرد.

موضوع‌گرا[۱۷]: یک انبارداده باید بتواند برای تحلیل یک حوزه‌ی موضوعی خاص مورد استفاده قرار بگیرد. برای مثال، «خدمات»، «متقاضیان»، «پیمان‌کاران»... می‌توانند هر کدام یک موضوع خاص باشند. روش‌های زیادی برای طبقه‌بندی یا موضوعی کردن داده‌ها وجود دارد اما چیزی که در پیاده‌سازی و به‌کارگیری یک انبارداده اهمیت دارد این است که تصمیم‌سازان به‌عنوان کاربران نهایی به دنبال تحلیل چه چیز هستند و منابع اطلاعاتی توسط چه بخش(هایی) تولید می‌شود. برای مثال دادگان استاندارد یک سازمان، ممکن است وظایفی را برای ذخیره‌سازی اطلاعات وام‌ها، کارت‌های اعتباری، پس‌انداز و

---

[14] Storage Layer
[15] Generation Layer
[16] Data Warehouse
[17] Subject-Oriented

ودیعه بر عهده داشته باشد (عملیات‌گرا) اما یک تحلیل‌گر بخواهد اطلاعاتی در مورد مشتریان، فروشندگان، محصولات و فعالیت آن‌ها داشته باشد. جهت‌بخشی به داده‌ها به‌سوی موضوعی خاص، منجر به انجام طبقه‌بندی‌هایی مفید برای یک تحلیل‌گر سازمانی خواهد شد.

یکپارچه[۱۸]: منابع داده‌ای متعدد درون یک انبارداده به‌صورت یک منبع اطلاعاتی یکپارچه در می‌آیند. برای مثال ممکن است منابع اطلاعاتی A و B حاوی روش‌های مختلفی برای شناسایی یک محصول یا ویژگی باشند. اما در یک انبارداده تنها یک روش برای شناسایی یک محصول یا ویژگی وجود خواهد داشت. به‌عبارتی دیگر خاصیت یکپارچگی یعنی با یک مدل واحد برای ذخیره‌سازی داده‌ها روبه‌رو هستیم. از آنجا که داده‌های عملیاتی یک سازمان در منابع گوناگونی ذخیره می‌شوند مدل داده‌ای هر کدام نیز متفاوت خواهد بود. هدف از پیاده‌سازی یک انبارداده این است که کاربر هنگام اجرای یک پرس و جو با مدل‌های داده‌ای مختلف سر و کار نداشته باشد. یکپارچه‌سازی داده‌ها در یک مکان و در یک مدل داده‌ای یکی از خواص اصلی یک انبارداده یا یکی از اهداف پیاده‌سازی آن است.

متغیر با زمان[۱۹]: داده‌های موجود در یک انبارداده مبتنی بر زمان نگهداری می‌شوند و می‌توان داده‌های مربوط به سه، شش، دوازده ماه پیش یا حتی قدیمی‌تر را از آن به‌دست آورد. این خاصیت در تضاد با سیستمی است که به دلایلی جدیدترین داده‌ها در آن نگهداری می‌شود. برای مثال یک سیستم ثبت معاملات ممکن است شامل آدرس‌های اخیر یک مشتری باشد در صورتی که یک انبارداده می‌تواند حاوی تمامی آدرس‌های مرتبط با آن مشتری باشد. داده‌های موجود در یک انبارداده می‌توانند در برگیرنده‌ی اطلاعات مربوط به یک دهه باشد در حالی که دادگان استاندارد معمولاً فقط شامل داده‌های یک ماه گذشته هستند. در یک انبارداده، زمان یک عامل کلیدی است چون به کاربر اجازه می‌دهد تا برای تحلیل خود یک روند در نظر داشته باشد. برای مثال با چنین

[18] Integrated
[19] Time-Variant

قابلیتی می‌توان یک ویژگی را برای یک فصل از سال استخراج و مورد بررسی قرار داد و آن را دوره‌های زمانی دیگر مقایسه کرد.

از دست نرفتنی[۲۰]: هنگامی که داده‌ای در انبارداده قرار می‌گیرد تغییر نخواهد کرد. بنا بر این داده‌های موجود در یک انبارداده که مبتنی بر تاریخ هستند هرگز نباید تغییر یابند. در یک دادگان استاندارد، داده‌ها می‌توانند اضافه، حذف یا روزامد شوند و معمولاً حالت فعلی سازمان را نشان می‌دهند. در یک انبارداده، داده‌های جدید فقط می‌توانند اضافه شوند و کاربران نمی‌توانند داده‌های موجود در آن را تغییر دهند. در انبارداده برخی سازمان‌ها حتی اطلاعات ۱۰ تا ۲۰ سال ذخیره می‌شود.

انبارداده موضوعی [۲۱]:یک انبارداده موضوعی انبارداده‌ی است که برای یک واحد ویژه در یک سازمان ایجاد شده است. برای مثال واحدهای مالی، فروش و بازایابی یک سازمان می‌توانند انبارداده موضوعی خود را داشته باشند. هم‌چنین تعریف دیگری برای انبارداده موضوعی به این صورت بیان می‌شود که یک انبارداده موضوعی یک لایه‌ی دسترسی به محیط انبارداده است و از آن برای استخراج داده برای کاربران استفاده می‌شود. به‌عبارت دیگر انبارداده موضوعی زیرمجموعه‌ای از یک انبارداده است که برای گروه مخصوصی در کسب و کار مورد نظر جهت‌دار شده است. عملاً انبارداده موضوعی قسمت‌های کوچکی از یک انبارداده هستند.

مرکز داده[۲۲]: مرکز داده، مجموعه‌ی بزرگی از سرورهای شبکه‌های رایانه‌ای است که معمولاً برای ذخیره‌سازی، پردازش یا توزیع حجم زیادی از داده‌ها از راه دور توسط سازمان‌ها مورد استفاده قرار می‌گیرد.

تلخیص[۲۳]: تلخیص، روشی برای افزایش سرعت اجرای پرس و جوها است. در این روش اطلاعات مرجع با توجه به ابعاد انتخاب‌شده توسط کاربر خلاصه می‌شوند.

استخراج؛ تبدیل و بارگذاری[۲۴]: فرایندی است که طی آن داده‌ها از یک محیط به محیط دیگر جابجا می‌شوند.

---

[20] Non-Volatile
[21] Data Mart
[22] Data Center
[23] Aggregation
[24] Extract, transform and load