

مفاهیم و ابزارهای داده‌کاوی

تألیف:

امید بزرگ جداد

استاد دانشگاه تهران

سهیلا زارعی

مهندسی منابع آب دانشگاه تهران



انتشارات
شهرک



انتشارات
شهرآب

مرکز بخش و دفتر مرکزی:
تهران، خیابان منبری جاوید (اردیبهشت)
کوچه بهشت آنین، پلاک ۳۶
کد پستی: ۱۳۱۴۹۶۳۳۳۱
تلفن: ۶۶۴۹۹۹۹۰
نمابر: ۶۶۴۸۳۱۵۱

فروشگاه مرکزی انتشارات شهرآب:
تهران، خیابان انقلاب،
روبروی دانشگاه تهران،
مجتمع فرهنگی فروزنده، پلاک ۲۳۳
تلفن: ۶۶۹۶۱۱۰۰ نمابر: ۶۶۹۶۱۱۰۱



انتشارات
آینده سازان

فروشگاه شماره ۲ انتشارات شهرآب:
تهران، خیابان انقلاب، ضلع جنوب
شرقی میدان انقلاب، بازار بزرگ کتاب
طبقه پایین همکف، پلاک ۱۵
تلفن: ۶۶۴۹۷۵۲۶

فروشگاه آینده سازان:
تهران، خیابان انقلاب، روبروی دانشگاه
تهران، مجتمع تجاری فروزنده، پلاک ۳۰۵
تلفن: ۶۶۴۱۷۳۱۶

مفاهیم و ابزارهای داده کاوی

تألیف : امید بزرگ حداد، سهیلا زارعی
ناشر : آینده سازان - شهرآب
نوبت چاپ : اول
تاریخ انتشار : ۱۴۰۰
تیراژ : ۱۱۰ نسخه
قیمت : ۱۵۰۰۰۰ تومان
چاپ و صحافی : ناژو
شابک : ۹۷۸-۹۶۴-۳۱۴-۴۷۷-۸

سرشناسه : بزرگ حداد، امید، ۱۳۵۲ -
عنوان و نام پدیدآور : مفاهیم و ابزارهای داده کاوی / تألیف امید بزرگ حداد، سهیلا زارعی.
مشخصات نشر : تهران : شهرآب: آینده سازان، ۱۴۰۰.
مشخصات ظاهری : ۲۴۸ ص: مصور (بخش رنگی)، عکس (رنگی)، نمودار.
شابک : 978-964-314-477-7
وضعیت فهرست نویسی : فیبا
پادداشت : کتابنامه.
موضوع : داده کاوی
Data mining
شناسه افزوده : زارعی، سهیلا، ۱۳۴۴ -
رده بندی کنگره : QA۷۶/۹
رده بندی دیویی : ۰۰۶/۳۱۲
شماره کتابشناسی ملی : ۸۵۴۶۵۵۹
اطلاعات رگورد کتابشناسی : فیبا

تمام حقوق نشر و بخش این اثر برای انتشارات شهرآب محفوظ است

پیشگفتار

روزانه هر شخص با تعداد زیادی از تصمیم‌گیری‌ها در مورد مسائل مختلف روزمره خود مواجه است. همچنین در امور حرفه‌ای، سازمان‌ها و مراکز مختلف نیز برای پیشبرد اهداف خود به تصمیم‌گیری نیاز دارند. به‌منظور اتخاذ بهترین تصمیم، این سازمان‌ها نیازمند مجموعه‌ای از داده‌های مختلف در ارتباط با هدف موردنظر خود هستند. مثلاً در دنیای اقتصاد، رفتار مشتری‌ها یکی از فاکتورهای تأثیرگذار در تصمیم‌گیری شرکت‌ها می‌باشد. زیرا از طریق آن می‌توان به عواملی که بر روی تصمیم‌گیری مشتری‌ها در خرید مؤثر است، پی برد. بنابراین شرکت‌ها باید آمار و اطلاعات آن را گردآوری و در پایگاه داده‌های خود نگهداری کنند. به این ترتیب در علوم مختلف و در سراسر جهان، حجم عظیمی از داده‌ها وجود دارند. از طرفی نیز شبکه جهانی وب، بشر را در دریای وسیعی از داده‌ها غوطه‌ور می‌کند و در عین حال، تصمیم‌های اتخاذ شده توسط هر فرد نیز ضبط می‌گردند.

در نتیجه افزایش روزانه حجم داده‌ها در بخش‌های مختلف زندگی، نگهداری این داده‌ها یکی از مسائل مهم پیش‌رو می‌باشد. برای این منظور مفهومی به نام پایگاه داده به وجود آمده است. پایگاه داده، جایی است که داده‌ها برای تجزیه و تحلیل‌های بعدی ذخیره و نگهداری می‌شوند که با ورود رایانه‌های شخصی در همه جا و دیسک‌های چندگیگابایتی، این مفهوم اهمیت، معنا و مفهوم پیدا کرده است. بنابراین پایگاه داده را می‌توان همانند مخزنی فرض کرد که داده‌ها را به‌صورت الکترونیکی و از طریق رایانه ذخیره و قابل دسترس می‌کند که معمولاً برای دستیابی به این داده‌ها روش‌های جست‌جو متفاوتی وجود دارد.

لازم به ذکر است که ذخیره و نگهداری داده‌های مختلف سودی نخواهد داشت. آن‌چه که مهم است اطلاعات و دانشی است که می‌توان از مجموعه داده‌های ذخیره شده کشف و استخراج کرد. زیرا در همه این داده‌ها، اطلاعات پنهان شده‌اند و هر دسته از این داده‌ها، از الگوهای مشخص و متفاوتی پیروی می‌کنند. به طور نمونه شکارچیان، الگوی رفتار مهاجرت حیوانات و یا کشاورزان، الگوی رشد گیاهان را می‌توانند جست‌جو کنند. در واقع اطلاعات مفیدی در این داده‌ها وجود دارند که به ندرت آشکار می‌شوند و یا مورد استفاده قرار می‌گیرند. مدت‌هاست اقتصاددان‌ها، آماردان‌ها

و مهندسان به دنبال راهکارهای خودکار برای جست‌جو، شناسایی، اعتبارسنجی و در نهایت استفاده از الگوهای موجود در داده‌ها برای مقاصد خود هستند. از طرف دیگر نیز این دانش استخراج شده باید در جایی نگهداری شود. این عملیات موجب به‌وجود آمدن مفهومی به نام پایگاه دانش شده است. پایگاه دانش، اطلاعات مفیدی است که از تجزیه و تحلیل داده‌های موجود در پایگاه داده به دست می‌آید.

تمام مطالب فوق نشان می‌دهند که فاصله فزاینده‌ای بین تولید داده‌ها و فهم انسان از آنها وجود دارد و همان‌طور که حجم داده‌ها افزایش می‌یابد، مقدار درک افراد از داده‌ها به‌طور اجتناب‌ناپذیر و نگران‌کننده‌ای کاهش می‌یابد. از طرف دیگر این مسئله، تصمیم‌گیری در مورد عملیاتی که باید بر روی تمام این داده‌ها اعمال گردند را نیز به تعویق انداخته‌اند. علم داده‌کاوی، تنها امید برای روشن شدن الگوهای پنهان موجود در داده‌ها است و پایه و اساس آن، همان داده‌هایی است که هر ساله حجم‌شان در حال افزایش است. پس اهمیت این علم از آن جهت است که می‌تواند با فرآیندهای محاسباتی به کشف الگوهای موجود در مجموعه‌ای بزرگ از داده‌ها کمک کند. بنابراین داده‌کاوی، فرآیند کشف الگوهای موجود در یک مجموعه از داده‌ها را به وسیله تحلیل داده‌ها برعهده دارد. الگوهای به دست آمده امکان انجام پیش‌بینی‌های مهمی براساس داده‌های جدید را فراهم می‌کنند. که همان نتیجه داده‌کاوی می‌باشد. در نتیجه می‌توان گفت که علم داده‌کاوی شامل دو مفهوم در ارتباط با داده‌ها، یعنی پایگاه داده و پایگاه دانش می‌باشد.

کتاب حاضر تلاشی در جهت گردآوری و ارائه مفاهیم و برخی فنون پرکاربرد داده‌کاوی به‌صورت جامع و یکپارچه است تا خواننده بتواند درک کاملی از این مفهوم با مقایسه روش‌های مختلف آن به دست آورد. تهیه کتابی با این مضمون از آن جهت حائز اهمیت است که با افزایش روزافزون حجم داده‌ها و به تبع آن حجم دانش و اطلاعات و همچنین کاربرد آنها در بخش‌ها و علوم مختلف، لازم است روش‌های کاربردی داده‌کاوی و تشکیل پایگاه دانش از پایگاه اطلاعات به علاقه‌مندان آموزش داده شود. لذا خواننده این کتاب با مهارت‌های تحلیل داده‌ها و فنون مختلف داده‌کاوی آشنا می‌شود.

فهرست مطالب

عنوان

شماره صفحه

فصل اول داده کاوی	۱
۱-۱- مقدمه	۱
۱-۲- تاریخچه	۳
۱-۳- تعریف داده کاوی	۵
۱-۴- عوامل مؤثر در توسعه داده کاوی و منشأ ایجاد داده کاوی	۹
۱-۵- وظایف داده کاوی	۱۲
۱-۶- داده کاوی و یادگیری ماشین	۱۵
۱-۷- جمع بندی	۱۹
۱-۸- منابع	۲۰
فصل دوم درخت تصمیم (DT)	۲۲
۲-۱- مقدمه	۲۲
۲-۲- تاریخچه	۲۳
۲-۳- ساختار و ویژگی های DT	۲۵
۲-۴- مفهوم DT و انواع آن	۳۰
۲-۴-۱- DT تک متغیره	۳۱
۲-۴-۲- DT چند متغیره	۳۱

- ۳۱.....DT دودویی ۳-۴-۲
- ۳۲.....۲-۵-۲ ویژگی‌های ورودی به گره
- ۳۲.....۲-۵-۱ ویژگی‌های وابسته به نام
- ۳۲.....۲-۵-۲ ویژگی‌های ترتیبی
- ۳۳.....۲-۵-۳ ویژگی‌های پیوسته
- ۳۵.....۶-۲ معیارهای ناخالصی جهت شناسایی ویژگی ورودی به گره
- ۳۶.....۲-۶-۱ معیار ناخالصی آنتروپی
- ۳۸.....۲-۶-۲ معیار ناخالصی جینی
- ۳۹.....۲-۶-۱-۲ نمایه جینی و DT دودویی
- ۴۰.....۲-۶-۲-۲ نمایه جینی و ورودی پیوسته
- ۴۱.....۳-۶-۲ معیار ناخالصی خطای طبقه‌بندی نادرست
- ۴۳.....۷-۲ آموزش DT
- ۴۵.....۲-۷-۱ الگوریتم هانت
- ۴۶.....۲-۷-۲ استفاده از تابع بازگشتی در آموزش DT
- ۴۷.....۸-۲ سود اطلاعات و نرخ سود
- ۵۰.....۹-۲ توقف رشد درخت و تولید برگ
- ۵۱.....۱۰-۲ مرز تصمیم‌گیری
- ۵۲.....۱۱-۲ درخت رگرسیون و تفاوت آن با درخت طبقه‌بندی
- ۵۴.....۲-۱۲-۱ DT در متلب
- ۵۴.....۲-۱۲-۱ انواع معیارهای بهینه‌سازی در متلب
- ۵۶.....۲-۱۲-۲ رشد بهینه DT در متلب

- ۵۷..... ۲-۱۳- دستورات پایه در متلب و حل مسأله
- ۶۱..... ۲-۱۴- جمع‌بندی
- ۶۱..... ۲-۱۵- منابع
- ۶۵..... فصل سوم شبکه عصبی مصنوعی (ANN)
- ۶۵..... ۳-۱- مقدمه
- ۶۶..... ۳-۲- تاریخچه
- ۶۹..... ۳-۳- ساختار و عملکرد نرون
- ۷۱..... ۳-۴- ANN و انواع ساختارهای آن
- ۷۴..... ۳-۴-۱- سلول عصبی تک‌ورودی
- ۷۶..... ۳-۴-۲- سلول عصبی چندورودی
- ۷۸..... ۳-۵-۱- توابع محرک یا انتقال
- ۷۸..... ۳-۵-۱- تابع خطی
- ۷۹..... ۳-۵-۲- تابع محرک به شدت محدود
- ۸۰..... ۳-۵-۳- تابع محرک غیرخطی
- ۸۱..... ۳-۵-۴- تابع لگاریتم سیگموئید
- ۸۳..... ۳-۵-۵- تابع تانژانت سیگموئید
- ۸۶..... ۳-۶- فرآیند تنظیم وزن
- ۸۶..... ۳-۷- معماری ANN
- ۸۷..... ۳-۷-۱- شبکه‌های عصبی پیش‌خور
- ۸۷..... ۳-۷-۲- شبکه‌های عصبی پس‌خور
- ۸۸..... ۳-۸- شبکه عصبی پرسپترون