

به نام خدا



مؤسسه فرهنگی هنری
دیباگران تهران

انبار داده‌ها و داده‌کاوی

مؤلف

دکتر مهدی اسماعیلی

هرگونه چاپ و تکثیر از محتویات این کتاب بدون اجازه کتبی ناشر ممنوع است. متخلفان به موجب قانون حمایت حقوق مؤلفان، مصنفان و هنرمندان تحت پیگرد قانونی قرار می گیرند.

انبار داده ها و داده کاوی

مؤلف: مهدی اسماعیلی

ناشر: مؤسسه فرهنگی هنری دیباگران تهران
حروفچینی و صفحه آرایی: دیباگران تهران

طرح روی جلد: مجتبی حجازی

چاپ: درج عقیق

نوبت چاپ: چهارم

تاریخ نشر: ۱۳۹۸

تیراژ: ۵۰ جلد

قیمت: ۱۰۰۰۰۰۰۰ ریال

شابک: ۹۷۸-۶۰۰-۱۲۴-۴۳۹-۱

نشانی واحد فروش: تهران، میدان انقلاب،

خ کارگر جنوبی، روبروی پاساژ مهستان،

پلاک ۱۲۵۱

تلفن: ۲۲۰۸۵۱۱۱-۶۶۴۱۰۰۴۶

کد پستی: ۱۳۱۴۹۸۳۱۸۵

فروشگاههای اینترنتی:

www.mftbook.ir

www.dibagaran-tehran.com

سرشناسه: اسماعیلی، مهدی، ۱۳۵۰-

عنوان و نام پدید آور: انبار داده ها و داده کاوی / مولف: مهدی اسماعیلی.

مشخصات نشر: تهران- دیباگران تهران- ۱۳۹۶

مشخصات ظاهری: ۴۱۶ ص. مصور.

شابک: ۹۷۸-۶۰۰-۱۲۴-۴۳۹-۱

وضعیت فهرست نویسی: فیبا

یادداشت: کتابنامه: ص (۳۹۳)-۴۱۶

موضوع: داده کاوی

موضوع: data mining

موضوع: داده پردازی

موضوع: electronic data processing

موضوع: پایگاههای اطلاعاتی - مدیریت

موضوع: database management

رده بندی کنگره: ۱۳۹۵ الف ۵/۲ QA ۷۶۹

رده بندی دیویی: ۰۰۶/۳۱۲

شماره کتابشناسی ملی: ۴۲۱۷۷۷۶

نشانی اینستاگرام: Dibagaran_publishing

نشانی تلگرام: @mftbook

پست الکترونیکی: bookmarket@mftmail.com

فهرست مطالب

فصل اول: مقدمه‌ای بر داده‌کاوی ۱۱

۱-۱ مقدمه	۱۱
۱-۲ کشف دانش	۱۱
۱-۳ داده‌کاوی و حوزه‌های دیگر	۱۵
۱-۴ اجرای پروژه‌های داده‌کاوی و متدولوژی CRISP-DM	۱۸
۱-۵ انواع داده‌ها جهت داده‌کاوی	۲۰
۱-۶ داده‌کاوی	۲۲
۱-۶-۱ الگوهای مکرر و قوانین انجمنی	۲۴
۱-۶-۲ دسته‌بندی	۲۵
۱-۶-۳ خوشه‌بندی	۲۸
۱-۷ چالش‌های داده‌کاوی	۳۲
۱-۸ سازماندهی کتاب	۳۳
خلاصه فصل	۳۴
منابع	۳۵

فصل دوم: آماده‌سازی داده‌ها ۳۷

۲-۱ مقدمه	۳۷
۲-۲ انواع صفات خاصه	۳۸
۲-۲-۱ متغیرهای کمی	۳۸
۲-۲-۲ متغیرهای کیفی	۳۹
۲-۳ آمار	۴۰
۲-۳-۱ شاخص‌های مرکزی	۴۰
۲-۳-۲ شاخص‌های پراکندگی	۴۳
۲-۳-۳ کوواریانس و ضریب همبستگی	۴۴
۲-۳-۴ توزیع نرمال	۴۵
۲-۴ لزوم آماده‌سازی داده‌ها	۴۷
۲-۵ تکنیک‌های آماده‌سازی داده‌ها	۴۹
۲-۵-۱ پالایش داده‌ها	۴۹

۶۰	۲-۵-۲ جمع‌آوری و یکپارچگی داده‌ها
۶۱	۲-۵-۳ تغییر شکل داده‌ها
۶۳	۲-۵-۴ کاهش داده‌ها
۶۴	۲-۵-۴-۱ کاهش صفات خاصه (کاهش ابعاد)
۷۴	۲-۵-۴-۲ کاهش نمونه‌ها
۷۶	۲-۵-۴-۳ کاهش مقادیر یک صفت خاصه
۸۱	خلاصه فصل
۸۴	منابع

فصل سوم: انبارش داده‌ها ۸۷

۸۷	۳-۱ مقدمه
۸۸	۳-۲ انبار داده‌ها
۹۰	۳-۳ پردازش تحلیلی برخط در مقابل پردازش تراکنشی برخط
۹۳	۳-۴ چرا یک مخزن جدا برای انبار داده‌ها لازم است؟
۹۴	۳-۵ مدل داده‌های چندبُعدی
۹۴	۳-۵-۱ مکعب‌ها
۱۰۰	۳-۵-۲ سلسله‌مراتب مفهومی
۱۰۱	۳-۵-۳ شمای پایگاه داده چندبُعدی
۱۰۶	۳-۵-۴ عملیات OLAP در مدل داده چندبُعدی
۱۱۰	۳-۶ دیتامارت
۱۱۱	۳-۷ معماری انبار داده‌ها
۱۱۲	۳-۷-۱ طراحی و ساخت انبار داده‌ها
۱۱۴	۳-۷-۲ معماری انبار داده‌ها
۱۱۶	۳-۷-۳ ابزارهای پسخوان انبار داده‌ها و مخزن متاداده‌ها
۱۱۷	۳-۷-۴ انواع سرویس‌دهنده‌های OLAP
۱۱۸	۳-۸ محاسبه موثر مکعب‌ها
۱۲۲	۳-۹ پردازش موثر پرسش‌های OLAP
۱۲۳	۳-۱۰ شاخص‌بندی داده‌های OLAP
۱۲۴	۳-۱۱ چند راهبرد کلی برای محاسبه مکعب داده‌ها
۱۲۶	خلاصه فصل
۱۲۸	منابع

فصل چهارم: الگوهای مکرر و قوانین انجمنی ۱۳۱

۱۳۱	۴-۱ مقدمه
۱۳۱	۴-۲ تحلیل سبد خرید
۱۳۳	۴-۳ قوانین انجمنی
۱۳۶	۴-۴ تولید الگوهای مکرر

۱۳۷Apriori الگوریتم ۴-۴-۱
۱۴۳FP-Growth الگوریتم ۴-۴-۲
۱۵۰ ۴-۵ تولید قوانین انجمنی
۱۵۱ ۴-۶ پایگاه داده تراکنشی و قالب‌های متفاوت داده‌ها
۱۵۳ ۴-۷ مجموعه اقلام ماکسیمال و بسته
۱۵۶ خلاصه فصل
۱۵۶ منابع

فصل پنجم: مباحث پیشرفته در قوانین انجمنی ۱۵۹

۱۵۹ ۵-۱ مقدمه
۱۵۹ ۵-۲ ارزیابی قوانین انجمنی
۱۶۷ ۵-۳ نکاتی پیرامون معیار پشتیبان
۱۷۰ ۵-۴ الگوهای نامکرر
۱۷۲ ۵-۴-۱ الگوهای منفی و الگوهای همبسته منفی
۱۷۴ ۵-۴-۲ تکنیک‌هایی برای کاوش الگوهای نامکرر
۱۷۷ ۵-۵ انواع دیگری از قوانین انجمنی
۱۷۸ ۵-۵-۱ صفات خاصه گسسته و قوانین انجمنی
۱۸۰ ۵-۵-۲ قوانین انجمنی کمی (مقداری)
۱۸۲ ۵-۵-۳ سلسله مراتب مفهومی و قوانین انجمنی
۱۸۴ ۵-۶ کاوش الگوهای مکرر مبتنی بر محدودیت
۱۸۸ خلاصه فصل
۱۸۸ منابع

فصل ششم: دسته‌بندی: مفاهیم پایه ۱۹۱

۱۹۱ ۶-۱ مقدمه
۱۹۲ ۶-۲ مفاهیم پایه
۲۰۰ ۶-۳ اندازه‌گیری خطا و برچسب‌هایی با مقادیر پیوسته
۲۰۲ ۶-۴ مجموعه داده‌های آموزشی و آزمایشی
۲۰۲ ۶-۴-۱ تکنیک Holdout
۲۰۳ ۶-۴-۲ تکنیک اعتبارسنجی متقابل
۲۰۴ ۶-۴-۳ تکنیک Bootstrap
۲۰۵ ۶-۵ روش‌هایی برای مقایسه مدل‌ها
۲۰۶ ۶-۵-۱ تخمین فاصله‌ی اطمینان
۲۰۸ ۶-۵-۲ مقایسه عملکرد دو مدل
۲۰۹ ۶-۵-۳ روشی دیگر برای تخمین فاصله اطمینان
۲۱۰ ۶-۵-۴ منحنی ROC
۲۱۲ ۶-۶ استفاده از روش‌های تلفیقی

۲۱۳	۶-۶-۱ روش Bagging
۲۱۳	۶-۶-۲ روش Boosting
۲۱۵	خلاصه فصل
۲۱۶	منابع

فصل هفتم: روش‌های دسته‌بندی ۲۱۷

۲۱۷	۷-۱ مقدمه
۲۱۷	۷-۲ درخت‌های تصمیم
۲۲۱	۷-۲-۱ معیارهای انتخاب صفت خاصه
۲۳۱	۷-۲-۲ چند موضوع دیگر در مورد درختان تصمیم
۲۳۲	۷-۲-۳ چند الگوریتم درخت تصمیم
۲۳۳	۷-۳ دسته‌بندی با کمک قانون بیز
۲۳۷	۷-۴ شبکه‌های بیز
۲۴۰	۷-۵ دسته‌بندی مبتنی بر قواعد
۲۴۳	۷-۵-۱ الگوریتم‌های یادگیری قواعد
۲۴۶	۷-۵-۲ معیارهای سنجش قواعد
۲۴۷	۷-۵-۳ بهینه نمودن قوانین
۲۴۸	۷-۶ ماشین‌های بردار پشتیبان
۲۴۹	۷-۶-۱ تفکیک‌پذیری خطی
۲۵۲	۷-۶-۲ تفکیک‌پذیر غیرخطی
۲۵۴	۷-۷ دسته‌بندی بر اساس تشابه
۲۵۴	۷-۷-۱ روش k همسایه نزدیک
۲۵۶	۷-۸ رگرسیون
۲۵۷	۷-۸-۱ رگرسیون خطی
۲۵۹	۷-۸-۲ رگرسیون غیرخطی و دیگر رگرسیون‌ها
۲۶۰	۷-۹ روش‌های دیگر برای دسته‌بندی
۲۶۰	۷-۹-۱ شبکه‌های عصبی
۲۶۳	۷-۹-۲ الگوریتم‌های ژنتیک
۲۶۴	۷-۹-۳ مجموعه‌های فازی
۲۶۵	۷-۱۰ چند موضوع درباره روش‌های دسته‌بندی
۲۶۹	خلاصه فصل
۲۷۱	منابع

فصل هشتم: خوشه‌بندی ۲۷۳

۲۷۳	۸-۱ مقدمه
۲۷۴	۸-۲ اهمیت و انگیزه خوشه‌بندی
۲۷۸	۸-۳ محاسبه تشابه و عدم تشابه

۲۸۰	۸-۳-۱ معیارهای تشابه برای داده‌های پیوسته
۲۸۵	۸-۳-۲ معیارهای تشابه و داده‌های دومقداره
۲۸۹	۸-۳-۳ معیارهای تشابه و داده‌های گسسته چندمقداره
۲۹۰	۸-۳-۴ محاسبه تشابه و انواع صفات خاصه
۲۹۱	۸-۴ روش‌های خوشه‌بندی
۲۹۴	۸-۵ روش‌های خوشه‌بندی مبتنی بر افراز داده‌ها
۲۹۴	۸-۵-۱ الگوریتم k-Means
۲۹۷	۸-۵-۲ الگوریتم k-Medoids
۲۹۹	۸-۵-۳ الگوریتم‌های دیگر
۳۰۰	۸-۶ تکنیک‌های خوشه‌بندی سلسله‌مراتبی
۳۰۲	۸-۶-۱ معیارهای تشابه میان خوشه‌ها
۳۰۹	۸-۶-۲ الگوریتم BIRCH
۳۱۱	۸-۶-۳ الگوریتم CURE
۳۱۲	۸-۶-۴ الگوریتم ROCK
۳۱۳	۸-۶-۵ الگوریتم Chameleon
۳۱۶	۸-۶-۶ الگوریتم DIANA
۳۱۸	خلاصه فصل
۳۱۹	منابع

فصل نهم: مباحث پیشرفته در خوشه‌بندی ۳۲۱

۳۲۱	۹-۱ مقدمه
۳۲۱	۹-۲ تکنیک‌های مبتنی بر چگالی
۳۲۲	۹-۲-۱ الگوریتم DBSCAN
۳۲۴	۹-۲-۲ الگوریتم OPTICS
۳۲۸	۹-۲-۳ الگوریتم DENCLUE
۳۳۱	۹-۳ تکنیک‌های مبتنی بر گرید
۳۳۲	۹-۳-۱ الگوریتم STING
۳۳۳	۹-۳-۲ الگوریتم CLIQUE
۳۳۶	۹-۴ تکنیک‌های خوشه‌بندی مبتنی بر مدل
۳۳۷	۹-۴-۱ الگوریتم EM
۳۳۸	۹-۴-۲ خوشه‌بندی مفهومی
۳۴۲	۹-۵ خوشه‌بندی با کمک الگوهای مکرر
۳۴۴	۹-۶ استفاده از محدودیت‌ها در خوشه‌بندی
۳۴۶	۹-۷ ارزشیابی روش‌های خوشه‌بندی
۳۴۷	۹-۷-۱ بررسی ساختار داده‌ها
۳۴۸	۹-۷-۲ تعیین تعداد خوشه‌ها
۳۴۹	۹-۷-۳ سنجش کیفیت خوشه‌بندی

۳۵۲.....	خلاصه فصل
۳۵۴.....	منابع

فصل دهم: روندها و مرزهای تحقیق ۳۵۷

۳۵۷.....	۱۰-۱ مقدمه
۳۵۷.....	۱۰-۲ گونه‌های دیگری از داده‌کاوی
۳۵۸.....	۱۰-۲-۱ متن‌کاوی
۳۶۰.....	۱۰-۲-۲ وب‌کاوی
۳۶۳.....	۱۰-۲-۳ گراف‌کاوی
۳۶۴.....	۱۰-۲-۴ کاوش داده‌های فضایی (مکان‌محور)
۳۶۶.....	۱۰-۲-۵ کاوش داده‌های چندرسانه‌ای
۳۶۸.....	۱۰-۲-۶ کاوش توالی‌ها
۳۷۲.....	۱۰-۳ داده‌کاوی، جامعه و امنیت
۳۷۶.....	۱۰-۴ کاربردهای داده‌کاوی
۳۷۷.....	۱۰-۴-۱ داده‌کاوی و بانکداری
۳۷۸.....	۱۰-۴-۲ داده‌کاوی و خرده‌فروشی
۳۷۹.....	۱۰-۴-۳ داده‌کاوی و مخابرات
۳۸۰.....	۱۰-۴-۴ داده‌کاوی و بیوانفورماتیک
۳۸۱.....	۱۰-۴-۵ هوش تجاری و داده‌کاوی
۳۸۲.....	۱۰-۴-۶ داده‌کاوی در علوم و مهندسی
۳۸۵.....	۱۰-۵ ابزارهای داده‌کاوی
۳۸۹.....	خلاصه فصل
۳۸۹.....	منابع

منابع و مراجع ۳۹۳

مقدمه ناشر

خط مشی کیفیت انتشارات مؤسسه فرهنگی هنری دیباگران تهران در عرصه کتاب‌هایی است که بتواند خواسته‌های به روز جامعه فرهنگی و علمی کشور را تا حد امکان پوشش دهد.

حمد و سپاس ایزد منان را که با الطاف بیکران خود این توفیق را به ما ارزانی داشت تا بتوانیم در راه ارتقای دانش عمومی و فرهنگی این مرز و بوم در زمینه چاپ و نشر کتب علمی دانشگاهی، علوم پایه و به ویژه علوم کامپیوتر و انفورماتیک گام‌هایی هرچند کوچک برداشته و در انجام رسالتی که بر عهده داریم، مؤثر واقع شویم.

گسترده‌ی علوم و توسعه روزافزون آن، شرایطی را به وجود آورده که هر روز شاهد تحولات اساسی چشمگیری در سطح جهان هستیم. این گسترش و توسعه نیاز به منابع مختلف از جمله کتاب را به عنوان قدیمی‌ترین و راحت‌ترین راه دستیابی به اطلاعات و اطلاع‌رسانی، بیش از پیش روشن می‌نماید. در این راستا، واحد انتشارات مؤسسه فرهنگی هنری دیباگران تهران با همکاری جمعی از اساتید، مؤلفان، مترجمان، متخصصان، پژوهشگران، محققان و نیز پرسنل ورزیده و ماهر در زمینه امور نشر درصدد هستند تا با تلاش‌های مستمر خود برای رفع کمبودها و نیازهای موجود، منابعی پُر بار، معتبر و با کیفیت مناسب در اختیار علاقمندان قرار دهند.

کتابی که در دست دارید با همت آقای مهدی اسماعیلی و تلاش جمعی از همکاران انتشارات میسر گشته که شایسته است از یکایک این گرامیان تشکر و قدردانی کنیم.

کارشناسی و نظارت بر محتوا: زهره قزلباش

در خاتمه ضمن سپاسگزاری از شما دانش‌پژوه گرامی درخواست می‌نماید با مراجعه به آدرس dibagaran.mft.info (ارتباط با مشتری) فرم نظرسنجی را برای کتابی که در دست دارید تکمیل و ارسال نموده، انتشارات دیباگران تهران را که جلب رضایت و وفاداری مشتریان را هدف خود می‌داند، یاری فرمایید.

امیدواریم همواره بهتر از گذشته خدمات و محصولات خود را تقدیم حضورتان نماییم.

مدیر انتشارات

مؤسسه فرهنگی هنری دیباگران تهران
Publishing@mftmail.com

مقدمه مؤلف

نه تو مانی، نه اندوه و نه هیچ یک از مردم این آبادی به تن لحظه خود جامه اندوه میپوشان هرگز امروزه رشد چشمگیر داده‌ها در سازمان، مدیران را مجبور ساخته تا از ابزارهای مناسبی جهت تصمیم‌گیری‌های خود استفاده کنند. در چنین محیطی جهت تحلیل داده‌ها و در نهایت اتخاذ یک تصمیم مدیریتی مناسب، لازم است اطلاعات کلیه بخش‌ها جمع‌آوری و یکپارچه شوند. در این وضعیت انبار داده‌ها یک راه‌حل مناسب تلقی می‌شود. هدف اصلی یک انبار داده‌ها تسهیل در فرآیند تصمیم‌گیری و افزایش دانش افراد درگیر در این فرآیند است. اگرچه وجود انبار داده‌ها پیش‌نیاز تحلیل داده‌ها و همچنین داده‌کاوی نیست، ولی با داشتن یک انبار داده‌ها، عمل داده‌کاوی نیز بسیار آسان‌تر و مطمئن‌تر انجام خواهد شد. داده‌کاوی همچون هر کاوش و تحلیل دیگری، به دنبال گنجی از اطلاعات و دانش سودمند در میان اقیانوسی از مجموعه داده‌ها است و رشد بی‌رویه داده‌ها، اهمیت آن را دو چندان کرده است.

موضوع کتاب حاضر انبار داده‌ها و داده‌کاوی است و مطالب آن در ۱۰ فصل گردآوری و تنظیم شده است. فصل اول کتاب به معرفی مفاهیم پایه داده‌کاوی می‌پردازد. روش‌های آماده‌سازی داده‌ها موضوع فصل دوم کتاب است. بسیاری از صاحب‌نظران اهمیت پیش‌پردازش داده‌ها را بخش مهمی از فرآیند داده‌کاوی می‌دانند. فصل سوم کتاب به موضوع انبارش داده‌ها (ایجاد انبار داده‌ها) تخصیص داده شده است. در فصل‌های چهارم و پنجم، در مورد قوانین انجمنی بحث می‌شود. روش‌های دسته‌بندی موضوع فصل‌های ششم و هفتم کتاب را تشکیل می‌دهد. خوشه‌بندی و مباحث پیشرفته آن نیز به ترتیب موضوع‌های فصل‌های هشتم و نهم می‌باشد. فصل دهم کتاب به صورت خلاصه به موضوعاتی از قبیل متن‌کاوی، وب‌کاوی، گراف‌کاوی و کاوش در میان داده‌های چندرسانه‌ای و فضایی می‌پردازد. بخش پایانی هر فصل نیز به معرفی منابع مربوط به موضوعات همان فصل می‌پردازد.

مطالب این کتاب به گونه‌ای نوشته شده‌اند تا خوانندگان محترم بتوانند مفاهیم آن را به راحتی درک کنند. واژه‌های معادل در این کتاب پیشنهادی هستند و چه بسا برای برخی از اصطلاحات، برابری دیگر (و احیاناً بهتر) بتوان یافت. آنچه مسلم است این است که اگرچه گاهی تبدیل واژه‌ها آنچنان که باید و شاید انجام نشده است، اما در ادای جملات و بیان موضوع تلاش فراوان شده است تا خوانندگان گرامی با متنی مبهم و گیج‌کننده روبرو نشوند.

در اینجا لازم می‌دانم از همه اساتید و دانشجویان به خاطر راهنمایی‌های ارزشمندشان در حین نگارش این کتاب سپاسگزاری کنم. همچنین از مدیریت محترم انتشارات دیباگران نیز به خاطر آماده‌سازی، چاپ و پخش این کتاب تشکر می‌کنم. رهین محبت بی‌دریغ خانواده‌ام هستم که با فراهم‌سازی محیطی مناسب مرا یار نمودند. با وجود همه سعی و تلاشی که در تمام مراحل آماده‌سازی این کتاب انجام گرفته است، یقین دارم که عاری از اشتباه نیست، چرا که تنها مکتوب بی‌نقص همان معجزه جاوید قرآن کریم است. در آخر ضمن سپاسگزاری از همه کسانی که مرا یاری داده‌اند و با پذیرش مسئولیت هرگونه کاستی احتمالی، امیدوارم که این اندک مفید افتد.

مهدی اسماعیلی

Msxpi2@yahoo.com