



مباحث پیشرفته در داده کاوی

مؤلفان

دکتر محمدرضا کیوانپور

مهندس فرانک حسن زاده

مهندس محمد مرادی



نشر دانشگاهی کیان
Kian Publication

سخنی با خوانندگان

«سپس، به کاتبان و نویسندگان بنگر و بهترین آن‌ها را بر کارهای خود بگمار...
کاتبان و نویسندگانی برگزین که قدر خود را بشناسند، چون کسی که به قدر خود شناخت
ندارد، دیگران را هم نمی‌شناسد.»
«برگرفته از نامه‌ی ۵۳ نهج البلاغه به مالک اشتر»

اگرچه نوشتن و پرداختن زکات علم از توصیه‌های اکید بزرگان و گواه بر کرامت اهل دانش است، اما
امروزه پرورش این به انگیزه‌ها و اهداف نوشتن بیشتر جلوه می‌کند. بی‌شک این‌که چه کسی می‌نویسد
مهم نیست، اما این‌که چرا و به چه پشتوانه‌ای می‌نویسد، درخور تأمل است. ما معتقدیم که چاپ
روزافزون کتاب‌های به اصطلاح «زرد» که خالی از هرگونه نوآوری و بی‌توجه به استانداردهای
چاپ کتاب و نیازهای طبع است، حاصل تفکر بازاری مستولی بر جامعه‌ی نشر است. بی‌پرده
آن‌که عنوان پر زرق و برق دستاویز قرار دادن مضمون‌های نو با هدف فروش بالا و طویل کردن
سیاهه‌ی سابقه‌ی علمی، نمی‌تواند دلیل محکمی برای چاپ و نشر کتابی باشد که خواننده‌ی مشتاق
با صرف هزینه‌های نه چندان کم آن را به دست می‌گیرد؛ به امید آن که چیزی از آن بیاموزد.

باید پذیرفت که انگیزه‌ی نوشتن کم و محتوای نوشته نیست و بین این دو رابطه‌ای مستقیم
برقرار است. اگر انگیزه از نوشتن، تولید دانش و بی‌شک نویسنده از قلم بی‌محتوا و کم‌عمق
پرهیز می‌کند و اگر دغدغه‌ی دانش و فرهنگ زخم خورده در میان باشد، ناشر تنها به عنوان
پرطمطراق بسنده نمی‌کند.

و چقدر امروزه، فرهنگ و دانش این مرز بوم که گرفته است بی‌انگیزگی و زخم هوس است،
نیازمند ناشران و نویسندگانی است که نیت‌شان کمک به رشد دانش و آفرینش‌های فرهنگ جامعه است و
به راستی که التیامی بر این درد نیست مگر نویسندگانی که قدر خود را می‌دانند و خوب
می‌فهمند که کتاب، ابزار سودجویی‌های مغرضانه نیست و می‌کوشند تا خوب و از هرگونه شهوت
نام و رسم و ثروت تهی کنند.

انتشارات دانشگاهی کیان خود را بری از عیب و خطا نمی‌داند، اما همواره بیشتر از پیش
می‌کوشیم تا در راستای تولید علم و نشر کتاب‌های پرمحتوا، دست نویسندگانی که انگیزه‌ی پاک
دارند را فشرده و در کنارشان باشیم و از خداوند متعال می‌خواهیم که در این مسیر صعب و پرخطر
در سایه‌ی لطف و عنایت خود از آن‌چه به عهده‌ی ما نهاده شده، سربلند و پیروز بر آییم.

انتشارات دانشگاهی کیان

در طی دهه گذشته حجم اطلاعات ذخیره شده در بانک‌های اطلاعاتی به صورت بسیار چشمگیری افزایش یافته است. در ابتدا تصور بر این بود که داده‌های ذخیره شده در سامانه‌های اطلاعاتی تا زمان مشخصی قابل استفاده و مفید هستند و پس از گذشت مدت زمان خاصی ارزش کاربردی خود را از دست می‌دهند. این تفکر در واقع، مبتنی بر استفاده سنتی از داده‌های بانک‌های اطلاعاتی بود که بیشتر بر پایه برداری از داده‌های مشخص و محسوس ذخیره شده در بانک‌های اطلاعاتی تاکید داشت. اما واقعیت این است که در هر مجموعه داده‌ای بزرگ علاوه بر داده‌های معمولی قابل استفاده، مجموعه‌ای از داده‌های نامحسوس و اطلاعات مفید نیز وجود دارد. استخراج این اطلاعات نامحسوس، اما ارزشمند می‌تواند به سبب استفاده از سیستم‌سازی‌های مؤثر در سازمان‌های مختلف فراهم آورد. داده‌کاوی علم کشف و استخراج اطلاعات مفید از مجموعه‌های داده‌ای بزرگ است. موضوع داده‌کاوی مبحثی به نسبت جدید و منتهای دانش است که به حوزه‌های آمار، هوش مصنوعی، یادگیری ماشین، بانک اطلاعاتی و سایر موضوعات و شناسایی الگو مرتبط است. فرایند داده‌کاوی در ابتدا روی داده‌های متنی ذخیره شده در بانک‌های اطلاعاتی انجام شد. پژوهشگران پس از مشاهده نتایج مفید و ارزشمند داده‌کاوی به توسعه و بهبود روش‌های مختلف انجام این فرایند جهت کاوش انواع گوناگون داده‌ها پرداختند. براین اساس نسل جدیدی از روش‌ها و کاربردها برای داده‌کاوی به وجود آمد.

این کتاب بر نسل جدید روش‌ها و کاربردهای داده کاوی متمرکز است. اگرچه مباحث مربوط به نسل جدید روش‌ها و کاربردهای داده‌کاوی به صورت پراکنده در مجلات مختلف منتشر شده؛ اما فقدان یک منبع جامع در این خصوص از یک سو و نیاز محققان و دانشجویان رشته‌های مختلف تحصیلی به چنین مرجعی، مولفان را بر آن داشت تا به تدوین این کتاب بپردازند. در هر یک از فصل‌های این کتاب یکی از مهم‌ترین و پرکاربردترین حوزه‌های داده‌کاوی به عنوان روش‌های و چالش‌های کاوش داده‌ها در آن حوزه معرفی شده است. اگرچه در تدوین کتاب تلاش برای بیان فنی استفاده شده، اما نحوه بیان مطلب به گونه‌ای است که مطالب برای افراد در رشته‌های مختلف مفید و قابل استفاده باشد. در پایان هر فصل مهم‌ترین منابع مورد استفاده در تدوین آن فصل ذکر شده است، اما منابع مورد استفاده محدود به موارد ذکر شده نمی‌باشد. همچنین برای درک بهتر اصطلاحات تخصصی واژه‌نامه‌ای کامل در پایان کتاب ارائه شده است. در پایان از همکاری دانشجویان و پژوهشگران محترم حامد حسن‌زاده، مهنوش خلقی، سهیلا مهرمولایی، مریم باحجب

ایمانی، سمیه سیفی مرادی، سعیده رنجبران، سمیه رحمانی، مریم پورکاظمی، مونا سلیمان پور، عاتکه اطاعتی، مریم کوهزادی، مهین رمضانی، صمیمانه در تدوین این کتاب مشارکت داشته‌اند تقدیر و تشکر به عمل می‌آید. همچنین از مدیریت محترم نشر دانشگاهی کیان که در مراحل مختلف تدوین کتاب یاریگر مولفان بودند تشکر و قدردانی به عمل می‌آید. از شما خوانندگان محترم تقاضا می‌شود نظرات سازنده خود را در راستای رفع نقایص کتاب و افزایش غنای محتوایی آن از طریق آدرس پست الکترونیکی advanced.datamining@gmail.com به مولفان ارسال نمایید.

عزیزان ما را یافید یافتید شاگردان مدرسه عشق را یاد کنید که همه ما
مدیون مردانگی آنها هستیم.

www.ketab.ir

فهرست مطالب

فصل اول: داده‌کاوی سنتی

۱۳	 مقدمه
۱۴	 ۱-۱. فرایند کشف دانش
۱۷	 ۲-۱. معرفی فرایند داده‌کاوی
۲۱	 ۳-۱. معرفی فرایند داده‌کاوی
۲۴	 ۴-۱. مراحل فرایند داده‌کاوی
۲۹	 ۵-۱. معماری یک سیستم داده‌کاوی
۳۲	 ۶-۱. چالش‌های داده‌کاوی
۳۸	 ۷-۱. انواع مختلف داده‌کاوی
۴۱	 ۸-۱. روش‌های داده‌کاوی
۴۴	 خلاصه فصل
۴۵	 منابع فصل اول

فصل دوم: متن‌کاوی

۴۹	 مقدمه
۵۰	 ۱-۲. تاریخچه و تعاریف اولیه متن‌کاوی
۵۸	 ۲-۲. چالش‌های کلی متن‌کاوی
۶۰	 ۳-۲. کاربردهای متن‌کاوی
۷۱	 ۴-۲. معماری سیستم‌های متن‌کاوی
۷۹	 ۵-۲. روش‌های متن‌کاوی
۸۴	 ۶-۲. رویکردهای آینده
۸۵	 خلاصه فصل
۸۶	 منابع فصل دوم

فصل سوم: وب‌کاوی

۸۹	 مقدمه
۹۰	 ۱-۳. وب: اصول، جنبه‌ها و رویکردها
۱۳۳	 ۲-۳. وب‌کاوی
۱۴۱	 ۳-۳. کاوش وب معنایی (وب‌کاوی معنایی)
۱۴۶	 ۴-۳. حوزه‌های کاربردی وب‌کاوی

۱۵۱	۳-۵. رویکردهای آینده
۱۵۱	خلاصه فصل
۱۵۲	منابع فصل سوم

فصل چهارم: کاوش شبکه‌های اجتماعی

۱۵۷	مقدمه
۱۵۸	۴-۱. شبکه‌های اجتماعی در یک نگاه کلی
۱۶۶	۴-۲. مرز شبکه‌های اجتماعی (SNM)
۱۷۳	۴-۳. چالش‌های کاوش شبکه‌های اجتماعی
۱۷۷	۴-۴. روش‌های کاوش و تحلیل شبکه‌های اجتماعی
۱۸۳	۴-۵. رویکردهای آینده
۱۸۳	خلاصه فصل
۱۸۴	منابع فصل چهارم

فصل پنجم: کاوش داده‌های بزرگ

۱۸۹	مقدمه
۱۹۰	۵-۱. داده‌های بزرگ
۱۹۸	۵-۲. کاوش داده‌های بزرگ
۲۰۰	۵-۳. تعریف جریان داده و انواع آن
۲۰۹	۵-۴. کاوش جریان‌های داده‌ای
۲۱۸	۵-۵. داده‌کاوی موازی و توزیعی (توزیع شده)
۲۲۹	۵-۶. رویکردهای آینده
۲۲۹	خلاصه فصل
۲۳۱	منابع فصل پنجم

فصل ششم: کاوش اسناد XML

۲۳۷	مقدمه
۲۳۸	۶-۱. مقدمه‌ای بر XML
۲۴۵	۶-۲. کاوش اسناد XML
۲۴۸	۶-۳. انواع اسناد XML
۲۵۲	۶-۴. چالش‌های کاوش اسناد XML
۲۵۳	۶-۵. روش‌های کاوش اسناد XML
۲۷۴	۶-۶. رویکردهای آینده

۲۷۶.....	خلاصه فصل.....
۲۷۷.....	منابع فصل ششم.....

فصل هفتم: کاوش داده‌های چندرسانه‌ای

۲۸۱.....	مقدمه.....
۲۸۲.....	۱-۷. مقدمه‌ای بر کاوش داده‌های چندرسانه‌ای.....
۲۹۷.....	۲-۷. کاوش داده‌های ویدیویی.....
۳۰۸.....	۳-۷. کاوش داده‌های صوتی (صوت‌کاوی).....
۳۱۴.....	۴-۷. صوت‌کاوی.....
۳۲۵.....	۵-۷. رویکردهای آینده.....
۳۲۶.....	خلاصه فصل.....
۳۲۷.....	منابع فصل هفتم.....

فصل هشتم: رخدادکاوی در داده‌های ویدیویی

۳۳۱.....	مقدمه.....
۳۳۲.....	۱-۸. معرفی و تعاریف اولیه.....
۳۴۰.....	۲-۸. چالش‌های رخدادکاوی.....
۳۴۴.....	۳-۸. کاربردهای رخدادکاوی.....
۳۴۹.....	۴-۸. رویکردهای موجود در حوزه رخدادکاوی.....
۳۵۶.....	۵-۸. دسته‌بندی روش‌ها و الگوریتم‌های رخدادکاوی.....
۳۷۲.....	۶-۸. رویکردهای آینده.....
۳۷۲.....	خلاصه فصل.....
۳۷۴.....	منابع فصل هشتم.....

فصل نهم: حرکت‌کاوی در داده‌های ویدیویی

۳۷۷.....	مقدمه.....
۳۷۸.....	۱-۹. اصول و مفاهیم حرکت‌کاوی در ویدیو.....
۳۸۳.....	۲-۹. کاربردهای حرکت‌کاوی در ویدیو.....
۳۸۷.....	۳-۹. رویکردهای حرکت‌کاوی در داده‌های ویدیویی.....
۳۹۵.....	۴-۹. تحلیل حرکت در ویدیو.....
۳۹۹.....	۵-۹. رویکردهای آینده.....
۴۰۰.....	خلاصه فصل.....
۴۰۱.....	منابع فصل نهم.....

فصل دهم: کاوش سری‌های زمانی

۴۰۳	مقدمه
۴۰۴	۱-۱۰. مقدمه‌ای بر سری‌های زمانی
۴۰۸	۲-۱۰. کاوش سری‌های زمانی
۴۱۳	۳-۱۰. حوزه‌های کاربردی کاوش سری‌های زمانی
۴۱۹	۴-۱۰. چالش‌های کلی در کاوش سری‌های زمانی
۴۲۰	۵-۱۰. رویکردهای کاوش سری‌های زمانی
۴۳۳	۶-۱۰. رویکردهای آینده
۴۳۴	خلاصه فصل
۴۳۶	منابع فصل دهم

فصل یازدهم: کاوش داده‌های فضایی - زمانی

۴۴۱	مقدمه
۴۴۲	۱-۱۱. داده‌های فضایی
۴۴۹	۲-۱۱. کاوش داده‌های فضایی
۴۵۸	۳-۱۱. داده‌های زمانی
۴۷۲	۴-۱۱. کاوش داده‌های زمانی (داده‌های فضایی)
۴۸۴	۵-۱۱. داده‌کاوی فضایی و زمانی (SD)
۴۸۷	۶-۱۱. رویکردهای آینده
۴۸۸	خلاصه فصل
۴۸۹	منابع فصل یازدهم

فصل دوازدهم: داده‌کاوی با حفظ حریم خصوصی

۴۹۳	مقدمه
۴۹۴	۱-۱۲. معرفی PPDM
۵۰۰	۲-۱۲. داده‌کاوی با حفظ حریم خصوصی چیست؟
۵۱۴	۳-۱۲. تشریح روش‌های داده‌کاوی با حفظ حریم خصوصی
۵۳۰	۴-۱۲. معیارهای ارزیابی روش‌های PPDM
۵۳۴	۵-۱۲. داده‌کاوی با حفظ حریم خصوصی مبتنی بر روش‌های رمزنگاری
۵۴۲	۶-۱۲. رویکردهای آینده
۵۴۲	خلاصه فصل
۵۴۴	منابع فصل دوازدهم

فصل سیزدهم: بازنمایی بصری (دیداری) و کاوش داده‌های بصری

مقدمه	۵۴۹
۱-۱۳. مقدمه‌ای بر بازنمایی داده‌ها	۵۵۰
۲-۱۳. بازنمایی بصری داده‌ها	۵۵۱
۳-۱۳. کاربردهای بازنمایی بصری	۵۶۱
۴-۱۳. روش‌های بازنمایی بصری	۵۶۳
۵-۱۳. کاوش بصری	۵۶۷
۶-۱۳. ریک‌های آینده	۵۷۲
خلاصه فصل	۵۷۲
منابع فصل سیزدهم	۵۷۳
واژه‌نامه	۵۷۵