بسم الله الرحمن الرحیم

دانشگاه تربیت دبیر شهید رجائی

پاک‌ســـازی داده‌هـــا

رفع ناسازگاری، تکرار، مقدار ازدست‌رفته و اغتشاش

تألیف:

دکتر نگین دانشپور

عضو هیأت علمی دانشگاه تربیت دبیر شهید رجائی

مهدیه عطاییان

سرشناسه	:	دانشپور، نگین، ۱۳۵۵-
عنوان و نام پدیدآور	:	پاک‌سازی داده‌ها: رفع ناسازگاری، تکرار، مقدار از دست‌رفته و اغتشاش/ تالیف نگین دانشپور، مهدیه عطاییان؛ ویراستار ادبی ساغر سلمانی‌نژاد.
مشخصات نشر	:	تهران: دانشگاه تربیت دبیر شهید رجایی، ۱۴۰۰.
مشخصات ظاهری	:	۲۲۰ ص.: جدول.
شابک	:	۹۷۸-۶۲۲-۶۵۸۹-۲۰-۸
وضعیت فهرست نویسی	:	فیپا
یادداشت	:	واژه‌نامه.
یادداشت	:	کتابنامه: ص. [۱۸۶] - ۱۸۸.
یادداشت	:	نمایه.
موضوع	:	داده‌پردازی
موضوع	:	Electronic data processing
موضوع	:	داده‌کاوی
موضوع	:	Data mining
شناسه افزوده	:	عطاییان، مهدیه، ۱۳۷۱-
شناسه افزوده	:	دانشگاه تربیت دبیر شهید رجایی
شناسه افزوده	:	Shahid Rajaee Teacher Training University
رده بندی کنگره	:	QA۷۶
رده بندی دیویی	:	۰۰۴
شماره کتابشناسی ملی	:	۸۵۴۱۹۸۸
اطلاعات رکورد کتابشناسی	:	فیپا

عنوان	:	پاک‌سازی داده‌ها: رفع ناسازگاری، تکرار، مقدار از دست رفته و اغتشاش
تألیف	:	دکتر نگین دانشپور، عضو هیأت علمی دانشگاه تربیت دبیر شهید رجائی/ مهدیه عطاییان
ویراستار ادبی	:	دکتر ساغر سلمانی‌نژاد
نوبت چاپ	:	اول- زمستان ۱۴۰۰
انتشارات	:	دانشگاه تربیت دبیر شهید رجائی
لیتوگرافی، چاپ	:	رجاء نقشینه، شریف
طراح جلد	:	محمد معتمدی‌نژاد
ناظر چاپ	:	محمد معتمدی‌نژاد
صفحه‌آرا	:	نیره فیروزی
شمارگان	:	۱۰۰ جلد
قیمت	:	۶۰۰.۰۰۰ ریال
شابک	:	۹۷۸-۶۲۲-۶۵۸۹-۲۰-۸ ISBN: 978-622-6589-20-8

نشانی: تهران، لویزان، کد پستی ۱۶۷۸۸-۱۵۸۱۱، صندوق پستی ۱۶۳ – ۱۶۷۸۵، تلفن: ۹ (۲۶۳۲ – ۲۲۹۷۰۰۶۰، ۲۲۹۷۰۰۷۰، تلفکس: ۲۲۹۷۰۰۴۲، پست الکترونیکی: publish@sru.ac.ir، وب سایت: http://publish.sru.ac.ir

پیش‌گفتار

داده‌های جمع‌آوری‌شده از منابع مختلف و توزیع‌شده برای اتخاذ تصمیمات مفید و سودمند باید به اطلاعات و دانش تبدیل شوند. به‌طور سنتی استخراج دانش را تحلیل‌گران داده انجام می‌دهند اما با توجه به رشد روزافزون داده‌ها، نیازمند روش‌هایی مبتنی بر رایانه هستیم. فرآیندهای مبتنی بر رایانه برای کشف دانش را «داده‌کاوی» می‌گویند. کیفیت داده‌های مورد بررسی در هنگام استخراج دانش، اهمیت بسزایی دارد. صحت، کامل‌بودن و سازگاری از جمله معیارهای مورد بررسی در بحث کیفیت داده‌ها هستند. داده‌های دنیای واقعی ممکن است به دلایل مختلفی دچار مشکل عدم کیفیت شوند. مقادیر جاافتاده، داده‌های مغشوش، ناسازگاری و مقادیر تکراری از جمله مشکلات عمده داده‌ها هستند. از این‌رو پیش‌پردازش که اغلب به منظور رفع این مشکلات انجام می‌شود، یکی از مهم‌ترین مراحل کشف دانش است.

پیش‌پردازش فرایندی زمان‌بر است اما با توجه به هزینه‌هایی که داده‌های فاقد کیفیت برای سازمان‌ها ایجاد می‌کنند، فرآیندی اجتناب‌ناپذیر است. درصورتی‌که دانش مورد نیاز برای اتخاذ تصمیمات، از داده‌های فاقد کیفیت استخراج شود، منجر به شکست تصمیمات می‌شود. پاک‌سازی و یکپارچه‌سازی از مراحل پیش‌پردازش هستند. در این کتاب تمرکز بر روی روش‌های جدید پاک‌سازی داده‌ها است. روش‌های مورد بررسی بر روی داده‌های عددی، ترتیبی و اسمی متمرکز هستند. در سال‌های اخیر پاک‌سازی داده‌ها از اهمیت ویژه‌ای برخوردار شده است، زیرا الگوریتم‌های داده‌کاوی برای عملکرد صحیح و ارائه دانش مفید، پیش‌بینی‌ها یا توصیف‌ها، نیاز به داده‌های صحیح، کامل و سازگار دارند. در فرایند پاک‌سازی داده‌ها هدف، تشخیص داده‌های مغشوش، تکراری، پرت و مقادیر جاافتاده است.

در این کتاب به معرفی مفاهیم اولیه پیش‌پردازش داده‌ها و ابعاد مختلف آن می‌پردازیم. همچنین مشکلات کیفیت داده‌ها به صورت اجمالی بیان می‌شوند و راهکارهای ابتدایی برای پاک‌سازی داده‌ها به‌طور مختصر معرفی می‌شوند. تمرکز اصلی در این کتاب بر روی شرح الگوریتم‌های پیشنهادی نویسندگان در حوزه پاک‌سازی داده‌ها است. کارایی این الگوریتم‌ها به دقت ارزیابی شده و برتری آن‌ها در مقایسه با کارهای معتبر جدید و قابل قیاس در آن حوزه تأیید شده است.

به این منظور بر روی این الگوریتم‌ها آزمایش‌های متعدد در شرایط مختلف انجام داده‌ایم. بر اساس آزمایش‌ها و ارزیابی‌های انجام‌شده، این نتیجه حاصل شد که این الگوریتم‌ها نسبت به الگوریتم‌های مشابه عملکرد بهتری دارند و بر این اساس در این کتاب ارائه کردیم.

با توجه به این‌که پاک‌سازی داده‌ها امری اجتناب‌ناپذیر در تحلیل داده‌ها است و در سال‌های اخیر این امر از اهمیت ویژه‌ای برخوردار شده است، هدف اصلی از ویرایش این کتاب ارائه یک نسخه فارسی متمرکز در این حوزه است. اگرچه منابع انگلیسی متعددی در زمینه پیش‌پردازش و کیفیت داده‌ها موجود است، اما منبع فارسی متمرکزی در حوزه پاک‌سازی داده‌ها موجود نیست. این کتاب می‌تواند برای محققان یا تحلیلگران داده در دانشگاه‌ها و مراکز تحقیقاتی، مفید واقع شود.

فهرست مطالب

فصل ۱

پیش‌پردازش داده‌ها

مقدّمه

رشد روزافزون داده‌ها موجب شده تا سازمان‌ها با داده‌های حجیم، منابع ناهمگون[1] و توزیع‌شده[2] روبرو شوند. استخراج دانش از داده‌ها برای ارائه خدمات و هدایت فرایند تصمیم‌گیری، به یک وظیفه اصلی در مدیریت داده‌ها تبدیل شده است [۱]. سازمان‌ها برای استخراج دانش و تصمیم‌گیری مبتنی بر داده‌های ذخیره‌شده یا گذرای[3] خود، نیازمند روش‌ها و ابزارهای خودکار هستند. وظیفه اصلی این ابزارها تبدیل حجم عظیم داده‌های خام به دانش و اطلاعات مفید است. این امر منجر به شکوفایی و رشد سریع دانش داده‌کاوی و کاربردهای مختلف آن در علوم رایانه در سال‌های اخیر شده است.

جمع‌آوری این حجم از داده‌ها و تبدیل آن‌ها به دانش، موجب ایجاد موضوعی به نام «کیفیت داده‌ها»[4] برای سازمان‌ها می‌شود. زمانی‌که داده‌ها از منابع و سامانه‌های مختلف به یک سامانه دیگر منتقل‌می‌شوند، ممکن است دچار مشکلاتی از قبیل قالب و دامنه ناهمگون، ناسازگاری[5] و مقادیر ازدست‌رفته[6] شوند. تصمیم‌گیری بر مبنای این داده‌های فاقد کیفیت، علاوه بر این‌که به ساختار سازمان آسیب می‌زند، باعث می‌شود هزینه‌های زیادی به سازمان وارد شود. بر اساس مطالعات انجام‌شده بیش از ۳۰٪ داده‌ها فاقد کیفیت هستند. این‌گونه داده‌ها موجب شده سالانه ۳تریلیون دلار به دولت آمریکا خسارت وارد شود [۳]؛ لذا تامین کیفیت داده‌ها در منابع داده از اهمیت بالایی برخوردار است.

کیفیت داده‌ها شامل شاخص‌های صحت[7]، کامل بودن[8]، سازگاری[9]، به هنگام‌بودن[10]، قابل اعتمادبودن[11] و تفسیرپذیری[12] است [۱۰]. داده‌ها در دنیای واقعی به دلایلی از جمله معیوب‌بودن تجهیزات جمع‌آوری اطلاعات، خطای سامانه و یا انسانی در هنگام واردکردن داده و قالب‌های ناهمگون در هنگام جمع‌آوری اطلاعات از منابع توزیع‌شده دچار مشکل عدم کیفیت می‌شوند. از

---

[1] Heterogeneous
[2] Distributed
[3] Transient
[4] Data quality
[5] Inconsistency
[6] Missing value
[7] Accuracy
[8] Completeness
[9] Consistency
[10] Timeliness
[11] Believability
[12] Interpretability

جمله متداول‌ترین مشکلات کیفیت داده‌ها می‌توان به مقادیر ازدست‌رفته، ناسازگاری، مقادیر ناصحیح[۱] و تکراری[۲] اشاره کرد.

پیش‌پردازش داده‌ها یکی از مهم‌ترین و ابتدایی‌ترین مراحل کشف دانش است که نقش اساسی در حصول اطمینان از کیفیت داده‌ها دارد. پاک‌سازی[۳] داده‌ها و یکپارچه‌سازی[۴] آن‌ها از جمله فرایندهایی هستند که در پیش‌پردازش انجام می‌گیرد. فراهم‌کردن داده‌هایی با کیفیت بالا کاری زمان‌بر و پرهزینه است؛ تا جایی‌که برآورد شده است که ۳۰٪-۸۰٪ هزینه و زمان ساخت پایگاه داده تحلیلی[۵] مربوط به پاک‌سازی داده است؛ اما صرف این هزینه و زمان زیاد در قبال هزینه‌های شکست بر اثر داده‌های فاقد کیفیت بسیار ارزشمند است [۳]. تضمین کیفیت داده‌ها در ساخت پایگاه داده تحلیلی بسیار حیاتی است.

بر این اساس در این فصل، در بخش ۱.۱ به معرفی کیفیت داده‌ها و ابعاد آن و در بخش ۲.۱ به بررسی مشکلات کیفیت داده و راهکارهای پاکسازی آن‌ها می‌پردازیم.

۱.۱. کیفیت داده‌ها و ابعاد آن

تضمین کیفیت داده‌ها از اهمیت بالایی برای سازمان‌ها برخوردار است؛ بالاخص زمانی‌که تصمیمات آینده سازمان، مبتنی بر این داده‌ها انجام پذیرد. منابع داده‌ای سازمان‌ها ممکن است به صورت ساخت‌یافته[۶]، نیمه‌ساخت‌یافته[۷] و غیرساخت‌یافته[۸] باشند. منابع ساخت‌یافته دارای موجودیتی[۹] با ساختار ثابت و مشخص هستند که مجموعه‌داده‌های رابطه‌ای[۱۰] و پایگاه داده‌های تحلیلی از رایج‌ترین انواع آن هستند. در منابع نیمه‌ساخت‌یافته موجودیت‌های مربوط به یک کلاس دارای ویژگی‌های متفاوتی هستند که رایج‌ترین نوع این منابع، داده‌های ارائه‌شده در قالب XML است. داده‌های منابع غیرساخت‌یافته دارای هیچ ساختار یا سازماندهی از پیش تعیین شده‌ای نیستند. فایل‌های متنی[۱۱] نمونه متداول از این منابع هستند [۱۳].

یک منبع داده درصورتی دارای کیفیت است که طرح[۱۲] و داده‌های آن دارای کیفیت باشند. کیفیت یک منبع، بر اساس میزان نزدیکی آن به نمونه‌هایش در دنیای واقعی سنجیده می‌شود. برای اندازه‌گیری این نزدیکی، از شاخص‌های کیفیت استفاده می‌شود. چنانچه طراحی اولیه طرح

---

[1] Incorrect value
[2] Duplicate
[3] Cleaning
[4] Integration
[5] Data warehouse
[6] Structured
[7] Semi-structured
[8] Unstructured
[9] Entity
[10] Relational
[11] Text
[12] Schema