

داده کاوی

www.ketab.ir

تألیف: دکتر جواد حسین خانی

سرشناسه	:	حسین خانی، جواد، ۱۳۵۸-
عنوان و نام پدیدآور	:	داده کاوی / تألیف جواد حسین خانی.
مشخصات نشر	:	تهران: یکان، ۱۳۹۲.
مشخصات ظاهری	:	۲۷۲ ص.: نمودار.
شابک	:	۱۲۰۰۰۰ ریال 978-964-8870-73-2
وضعیت فهرست نویسی	:	فیا
موضوع	:	داده کاوی
رده بندی کنگره	:	۱۳۹۲ ج ۴۷/۹/۴۷ Q۸۷۷
رده بندی دیویی	:	۰۰۶/۳۱۲
شماره کتابشناسی ملی	:	۳۱۳۱۲۰

داده کاوی

تألیف: دکتر جواد حسین خانی

چاپ اول: بهار ۱۳۹۲

تیراژ: ۵۰۰ نسخه

ناشر: یکان

شابک: ۷۳-۲-۸۸۷۰-۸۸۷۰-۹۶۴-۹۷۸

قیمت: ۱۲۰۰۰ تومان

کلیه حقوق مادی و معنوی این اثر متعلق به مؤلف می باشد و استفاده از مطالب این اثر بدون مجوز رسمی و کتبی از انتشارات و یا مؤلف پیگرد قانونی دارد.

انتشارات یکان: خیابان انقلاب، خیابان منیری جاوید (اردیبهشت)، پلاک ۶۱، ساختمان دانشجو،

طبقه چهارم، واحد ۷ تلفن: ۶۶۴۱۵۶۸۱ - ۶۶۹۵۰۰۱۵

سخن نگارنده

علم نوین داده‌کاوی به طور روزافزون در حال توسعه است. بسیاری از شرکت‌ها و موسسات، دارای حجم انبوهی از اطلاعات هستند. تکنیک‌های داده‌کاوی به گونه‌ای گسترش یافته‌اند که می‌توان آنها را بر ابزارهای نرم‌افزاری امروزی تطبیق داده و از اطلاعات استنتاج شده بهترین استفاده را برد. این کتاب موضوعات اصلی داده‌کاوی و تشریح مفاهیم و تکنیک‌های آن، از جمله مدل‌های آماری و الگوریتم‌های ریاضی را پوشش می‌دهد و مناسب دانشجویان و محققین گرایش‌های مختلف مهندسی کامپیوتر، صنایع، ریاضی و آمار می‌باشد.

از خوانندگان گرامی به ویژه اساتید بزرگوار، متخصصین گرامی و دانشجویان عزیز تقاضا می‌شود با نظرات و پیشنهادات سازنده خود از طریق آدرس پست الکترونیکی jhkhani@gmail.com نگارنده را مورد لطف خود قرار دهند.

دکتر جواد حسین خانی

فروردین ۱۳۹۲

فهرست مطالب

صفحه	عنوان
۱۱	فصل اول: پیش‌گفتار
۱۱	۱- مقدمه
۱۳	۱-۱ داده‌کاوی
۱۷	۱-۲ محدودیت‌های داده‌کاوی
۱۸	۱-۳ خلاصه فصل‌ها
۲۱	۱-۴ نحوه مطالعه کتاب و توصیه به اساتید
۲۳	فصل دوم: قوانین انجمنی و الگوهای ترتیبی
۲۴	۲-۱ مفاهیم ابتدایی در قوانین انجمنی
۲۸	۲-۲ الگوریتم استقرایی
۲۸	۲-۲-۱ استخراج مجموعه موارد تکراری
۳۴	۲-۲-۲ تولید قوانین انجمنی
۳۶	۲-۳ قالب‌های داده‌ای برای کاوش قوانین انجمنی
۳۸	۲-۴ کاوش با حداقل پشتیبانی چندگانه
۴۱	۲-۴-۱ مدل توسعه یافته
۴۳	۲-۴-۲ الگوریتم کاوش
۵۱	۲-۴-۳ تولید قانون
۵۳	۲-۵ کاوش قوانین انجمنی کلاس
۵۳	۲-۵-۱ تعریف مسئله
۵۵	۲-۵-۲ الگوریتم کاوش
۵۹	۲-۵-۳ کاوش با حداقل پشتیبانی چندگانه
۶۰	۲-۶ مفاهیم پایه الگوهای ترتیبی
۶۳	۲-۷ کشف الگوهای ترتیبی بر پایه GSP
۶۳	۲-۷-۱ الگوریتم GSP
۶۵	۲-۷-۲ کاوش با حداقل پشتیبانی چندگانه

۷۱	۲-۸ کشف الگوهای ترتیبی بر اساس PrefixSpam
۷۲	۲-۸-۱ الگوریتم PrefixSpam
۷۴	۲-۸-۲ کاوش با حداقل پشتیبانی چندگانه
۷۷	۲-۹ تولید قواعد از الگوهای ترتیبی
۷۷	۲-۹-۱ قواعد ترتیبی
۷۸	۲-۹-۲ قوانین نشانه گذاری ترتیبی (LSR)
۷۹	۲-۹-۳ قواعد ترتیبی کلاس (CSR)
۸۰	یادداشتی بر فصل
۸۳	منابع و مآخذ
۸۵	فصل سوم: یادگیری نظارت شده
۸۶	۳-۱ مفاهیم اولیه
۹۱	۳-۲ استنتاج درخت تصمیم
۹۶	۳-۲-۱ الگوریتم یادگیری
۹۷	۳-۲-۲ تابع ناخالصی
۱۰۲	۳-۲-۳ مدیریت صفات پیوسته
۱۰۵	۳-۲-۴ چند مسأله دیگر
۱۰۹	۳-۳ ارزیابی طبقه‌بند
۱۰۹	۳-۳-۱ روش‌های ارزیابی
۱۱۱	۳-۳-۲ دقت، فراخوانی، F-score و Breakeven point
۱۱۵	۳-۴ استنتاج قانون
۱۱۵	۳-۴-۱ پوشش ترتیبی
۱۱۸	۳-۴-۲ آموختن قانون: تابع Learn-One-Rule
۱۲۳	۳-۴-۳ تشریح مطالب
۱۲۴	۳-۵ طبقه‌بندی بر اساس انجمن‌ها
۱۲۵	۳-۵-۱ طبقه‌بندی با استفاده از قوانین انجمنی کلاس
۱۳۱	۳-۵-۲ قوانین انجمنی کلاس به عنوان خصوصیت
۱۳۱	۳-۵-۳ طبقه‌بندی با استفاده از قوانین انجمنی نرمال
۱۳۳	۳-۶ طبقه‌بندی بیزی ساده

صفحه	عنوان
۱۳۸	۳-۷ طبقه‌بندی متن بیزی ساده
۱۳۹	۳-۷-۱ چارچوب احتمالی
۱۴۱	۳-۷-۲ مدل بیزی ساده
۱۴۵	۳-۷-۳ تشریح مطالب
۱۴۶	۳-۸ ماشین‌های برداری
۱۴۹	۳-۸-۱ ماشین‌های برداری خطی: با موارد قابل تفکیک
۱۵۷	۳-۸-۲ ماشین‌های برداری خطی: غیر قابل تفکیک
۱۶۲	۳-۸-۳ ماشین‌های برداری غیر خطی: توابع کرنل
۱۶۷	۳-۹ یادگیری نزدیک‌ترین مجاور
۱۷۰	۳-۱۰ اثر کلی کلاس‌بندی‌ها
۱۷۰	۳-۱۰-۱ مدل بسته‌ای
۱۷۱	۳-۱۰-۲ تکنیک بالارونده
۱۷۳	یادداشتی بر فصل
۱۷۶	منابع و مآخذ
۱۷۹	فصل چهارم: یادگیری بدون نظارت
۱۸۰	۴-۱ مفاهیم پایه
۱۸۳	۴-۲ خوشه‌بندی k-means
۱۸۴	۴-۲-۱
۱۸۷	۴-۲-۲ مدل دیسک الگوریتم k-means
۱۸۸	۴-۲-۳ نقاط ضعف و قوت
۱۹۴	۴-۳ نحوه نمایش خوشه‌ها
۱۹۴	۴-۳-۱ روش‌های معمول نمایش خوشه‌ها
۱۹۶	۴-۳-۲ خوشه‌های شکل‌های اختیاری
۱۹۷	۴-۴ خوشه‌بندی سلسله مراتبی
۱۹۹	۴-۴-۱ روش Single-Link
۲۰۰	۴-۴-۲ روش Complete-Link
۲۰۱	۴-۴-۳ روش Average-Link

صفحه	عنوان
۲۰۱	۴-۴-۴ ضعف و قدرت
۲۰۲	۴-۵ تابع‌های فاصله
۲۰۲	۴-۵-۱ صفات عددی
۲۰۴	۴-۵-۲ خواص اسمی و باینری
۲۰۶	۴-۵-۳ مستندات متنی
۲۰۷	۴-۶ استانداردسازی داده
۲۱۰	۴-۷ کار با خاصیت‌های مرکب
۲۱۳	۴-۸ کدام الگوریتم خوشه‌بندی استفاده شود؟
۲۱۳	۴-۹ ارزیابی خوشه
۲۱۷	۴-۱۰ پیدا کردن حفره‌ها و مناطق داده‌ها
۲۲۱	منابع و مآخذ
۲۲۳	فصل پنجم: یادگیری نظارتی به طور جزئی
۲۲۴	۵-۱ یادگیری از نمونه‌های برچسب‌گذاری شده و برچسب‌گذاری نشده
۲۲۶	۵-۱-۱ الگوریتم EM با طبقه‌بندی ساده Naive
۲۳۱	۵-۱-۲ هم آموزش
۲۳۵	۵-۱-۳ خود آموزشی
۲۳۶	۵-۱-۴ ماشین‌های بردار پشتیبانی انتقالی
۲۳۸	۵-۱-۵ روش‌های مبتنی بر گراف
۲۴۳	۵-۱-۶ بحث و گفتگو
۲۴۴	۵-۲ یادگیری از نمونه‌های مثبت و برچسب‌گذاری نشده
۲۴۵	۵-۲-۱ کاربردهای یادگیری PU
۲۴۹	۵-۲-۲ مبنای نظری
۲۵۱	۵-۲-۳ طبقه‌بندی‌کننده‌های ساختمان: روش دو گام
۲۶۰	۵-۲-۴ طبقه‌بندهای ساختمان: رویکرد مستقیم
۲۶۳	۵-۲-۵ بحث و گفتگو
۲۶۸	یادداشتی بر فصل
۲۷۰	منابع و مآخذ

فصل اول

پیش گفتار

۱- مقدمه

با توسعه سیستم‌های پایگاهی و حجم زیاد داده‌های ذخیره شده در این سیستم‌ها، به ابزاری نیاز داریم تا بتوان پس از پردازش و تحلیل داده‌ها، سطح بالاتری از اطلاعات موجود را در اختیار کاربران قرار داد. یکی از روش‌های بسیار مهم که با آن می‌توان به کشف الگوهای مفید در میان داده‌ها پرداخت، داده‌کاوی است. این روش که با حداقل دخالت کاربران همراه است اطلاعاتی را در اختیار آنها و تحلیل گران قرار می‌دهد تا بر اساس آنها تصمیمات مهم و حیاتی در سازمانشان اتخاذ نمایند. لازم به ذکر است که داده‌کاوی زمانی مطرح می‌شود که با حجم زیادی از داده‌ها مواجه باشیم. هرچه حجم داده‌ها بیشتر و روابط میان آنها پیچیده‌تر باشد دسترسی به اطلاعات نهفته در میان داده‌ها مشکلتر می‌شود و نقش داده‌کاوی به عنوان یکی از روش‌های کشف دانش، آشکارتر می‌شود. در واقع داده‌کاوی پایگاه‌ها و مجموعه‌های حجیم داده‌ها را در پی کشف و استخراج دانش، مورد تحلیل و کندوکاوهای اتوماتیک و نیمه اتوماتیک قرار می‌دهد.

بسیاری از شرکت‌ها و موسسات دارای حجم انبوهی از اطلاعات هستند. تکنیک‌های داده‌کاوی به طور تاریخی به گونه‌ای گسترش یافته‌اند که به سادگی می‌توان آنها را بر ابزارهای نرم‌افزاری امروزی تطبیق داده و از اطلاعات جمع‌آوری شده بهترین بهره را برد. در صورتی که سیستم‌های داده‌کاوی بر روی سکوهاى Client/Server قوی نصب شده باشند و دسترسی به بانک‌های اطلاعاتی بزرگ فراهم باشد، می‌توان به سؤالاتی از قبیل: کدامیک از مشتریان ممکن است خریدار کدامیک از محصولات آینده شرکت باشند، چرا، در کدام مقطع زمانی و بسیاری از موارد مشابه پاسخ داد.

یکی از ویژگی‌های کلیدی در بسیاری از ابتکارات مربوط به تأمین امنیت ملی، داده‌کاوی است. داده‌کاوی که به عنوان ابزاری برای کشف جرایم، ارزیابی میزان ریسک و فروش محصولات به کار می‌رود، در برگیرنده ابزارهای تجزیه و تحلیل اطلاعات به منظور کشف الگوهای معتبر و ناشناخته در بین انبوهی از داده‌هاست. داده‌کاوی غالباً در زمینه تأمین امنیت ملی به منزله ابزاری برای شناسایی فعالیت‌های افراد خرابکار شامل جابه‌جایی پول و ارتباطات بین آنها و همچنین شناسایی و ردگیری خود آنها با بررسی سوابق مربوط به مهاجرت و مسافرت‌هاست. داده‌کاوی پیشرفت قابل ملاحظه‌ای را در نوع ابزارهای تحلیل موجود نشان می‌دهد اما محدودیت‌هایی نیز دارد. یکی از این محدودیت‌ها این است که با وجود اینکه به آشکارسازی الگوها و روابط کمک می‌کند اما اطلاعاتی را درباره ارزش یا میزان اهمیت آنها به دست نمی‌دهد. دومین محدودیت آن این است که با وجود توانایی شناسایی روابط بین رفتارها و یا متغیرها لزوماً قادر به کشف روابط علت و معلولی نیست. موفقیت داده‌کاوی در گروه بهره‌گیری از کارشناسان فنی و تحلیل‌گران کار آزموده‌ای است که از توانایی کافی برای طبقه‌بندی تحلیل‌ها و تغییر آنها برخوردار هستند. بهره‌برداری از داده‌کاوی در دو بخش دولتی و خصوصی رو به گسترش است. صنایعی چون بانکداری، بیمه، بهداشت و بازاریابی آن را عموماً برای کاهش هزینه‌ها، ارتقاء کیفی پژوهش‌ها و بالاتر بردن میزان فروش به کار می‌برند. کاربرد اصلی داده‌کاوی در بخش دولتی به عنوان ابزاری برای تشخیص جرایم بوده‌است اما امروزه دامنه بهره‌برداری از آن

گسترش روزافزونی یافته و سنجش و بهینه‌سازی برنامه‌ها را نیز در برمی‌گیرد. بررسی برخی از برنامه‌های کاربردی مربوط به داده‌کاوی که برای تامین امنیت ملی به کار می‌روند، نشان‌دهنده رشد قابل ملاحظه‌ای در رابطه با کمیت و دامنه داده‌هایی است که باید تجزیه و تحلیل شوند. توانایی‌های فنی در داده‌کاوی از اهمیت ویژه‌ای برخوردار اند اما عوامل دیگری نیز مانند چگونگی پیاده‌سازی، نظارت ممکن است نتیجه کار را تحت تأثیر قرار دهند. یکی از این عوامل کیفیت داده‌هاست که بر میزان دقت و کامل بودن آن دلالت دارد. عامل دوم میزان سازگاری نرم‌افزار داده‌کاوی با بانکهای اطلاعاتی است که از سوی شرکت‌های متفاوتی عرضه می‌شوند. عامل سومی که باید به آن اشاره کرد به بیراهه رفتن داده‌کاوی و بهره‌برداری از داده‌ها به منظوری است که در ابتدا با این نیت گرد آوری نشده‌اند. حفظ حریم خصوصی افراد عامل دیگری است که باید به آن توجه داشت. اصولاً به پرسش‌های زیر در زمینه داده‌کاوی باید پاسخ داده شود:

- سازمانهای دولتی تا چه حدی مجاز به بهره‌برداری از داده‌ها هستند؟
- آیا از داده‌ها در چارچوبی غیر متعارف بهره‌برداری می‌شود؟
- کدام قوانین حفظ حریم خصوصی ممکن است به داده‌کاوی مربوط شوند؟
- کشف دانش در داده‌ها بخش بزرگی از سامانه‌های هوشمند است. سامانه‌های هوشمند زیرشاخه‌ای بزرگ و پرکاربرد از یادگیری ماشین در هوش مصنوعی می‌باشد.

۱-۱ داده‌کاوی^۱

به عنوان کشف دانش در بانک اطلاعاتی^۲ تلقی می‌گردد. داده‌کاوی، به طور کلی به عنوان فرآیند کشف الگوهای^۳ مفید یا دانش از منابع اطلاعاتی نظیر بانک‌های اطلاعاتی، متن‌ها، تصاویر، وب و غیره تعریف می‌شود. الگوها بایستی قابل قبول، بالقوه سودمند و قابل فهم

1. Data Mining

2. Knowledge Discovery in Database (KDD)

3. patterns

باشند. داده‌کاوی یک فیلد چندوجهی شامل یادگیری ماشین^۱، آمار^۲، بانک اطلاعاتی، هوش مصنوعی^۳، فراخوانی اطلاعات^۴ و بصری‌سازی^۵ محسوب می‌گردد. داده‌کاوی کاربردهای زیادی دارد. بعضی از متداول‌ترین آنها عبارتند از یادگیری نظارت شده^۶ (یا طبقه‌بندی^۷)، یادگیری بدون نظارت^۸ (یا دسته‌بندی^۹)، کاوش قوانین انجمنی^{۱۰} و کاوش الگوهای ترتیبی^{۱۱}. در این کتاب تمام این موارد مورد بررسی قرار می‌گیرند.

داده‌کاوی به بهره‌گیری از ابزارهای تجزیه و تحلیل داده‌ها به منظور کشف الگوها و روابط معتبری که تا کنون ناشناخته بوده‌اند اطلاق می‌شود. این ابزارها ممکن است مدل‌های آماری، الگوریتم‌های ریاضی و روش‌های یادگیرنده باشند که کار این خود را به صورت خودکار و بر اساس تجربه‌ای که از طریق شبکه‌های عصبی^{۱۲} یا درخت‌های تصمیم^{۱۳} به دست می‌آورند بهبود می‌بخشد. داده‌کاوی منحصر به گردآوری و مدیریت داده‌ها نبوده و تجزیه و تحلیل اطلاعات و پیش‌بینی را نیز شامل می‌شود برنامه‌های کاربردی که با بررسی فایل‌های متن یا چندرسانه‌ای به کاوش داده‌ها می‌پردازند پارامترهای گوناگونی را در نظر می‌گیرد که عبارت‌اند از:

- رابطه^{۱۴}: الگوهایی که بر اساس آن یک رویداد به دیگری مربوط می‌شود مثلاً خرید قلم به خرید کاغذ.

- ترتیب^{۱۵}: الگویی که به تجزیه و تحلیل توالی رویدادها پرداخته و مشخص می‌کند کدام رویداد، رویدادهای دیگری را در پی دارد مثلاً خرید تشک و خرید تخت‌خواب.

1. machine learning
2. statistics
3. Artificial Intelligence
4. Information Retrieval
5. Visualization
6. supervised learning
7. classification
8. unsupervised learning
9. clustering
10. Association Rule Mining
11. Sequential pattern Mining
12. Neural Networks
13. Decision Trees
14. Association
15. Sequence

- دسته‌بندی^۱: شناسایی الگوهای جدید مثلاً همزمانی خرید چسب و پوشه.

- خوشه‌بندی^۲: کشف و مستندسازی مجموعه‌ای از حقایق ناشناخته مثلاً موقعیت جغرافیایی خرید محصولی با مارک خاص.

- پیش‌بینی^۳: کشف الگوهایی که بر اساس آنها پیش‌بینی قابل قبولی از رویدادهای آتی ارایه می‌شود، مثلاً رابطه عضویت در یک باشگاه ورزشی با شرکت در کلاسهای ورزشی.

- مصورسازی^۴: مصورسازی داده‌ها یکی از قدرتمندترین و جذاب‌ترین روش‌های اکتشاف در داده‌ها می‌باشد.

برنامه‌های کاربردی که در زمینه تجزیه و تحلیل اطلاعات به کار می‌روند از امکاناتی چون پرس و جوی ساخت‌یافته^۵ که در بسیاری از بانک‌های اطلاعاتی یافت می‌شود و از ابزارهای تجزیه و تحلیل آماری برخوردارند اما برنامه‌های مربوط به داده‌کاوی در عین برخورداری از این قابلیت‌ها از نظر نوع با آنها تفاوت دارند.

در نتیجه قابلیت‌های پیچیده‌اش برای موفقیت در تمرین داده‌کاوی دو مقدمه مهم است یکی فرمول واضحی از مشکل که قابل حل باشد و دیگری دسترسی به داده متناسب. بعضی از ناظران داده‌کاوی را مرحله‌ای در روند کشف دانش در پایگاه داده‌ها می‌دانند. مراحل دیگری در روند کشف دانش به صورت تساعدی شامل، پاکسازی داده، انتخاب داده، انتقال داده، داده‌کاوی، الگوی ارزیابی، و عرضه دانش می‌باشد. بسیاری از اینترنت‌ها در تکنولوژی و فرایندهای تجاری بر رشد علاقه‌مندی به داده‌کاوی در بخش‌های خصوصی و عمومی سهمی داشته‌اند. بعضی از این تغییرات شامل:

- رشد شبکه‌های کامپیوتری که در ارتباط برقرار کردن پایگاه‌ها داده مورد استفاده قرار

می‌گیرند.

1. Classification
2. Clustering
3. Forecasting
4. visualization
5. Structured query

- توسعه افزایش تکنیک‌هایی بر پایه جستجو مثل شبکه‌های عصبی و الگوریتم‌های پیشرفته.

- گسترش مدل محاسبه کلاینت سروری که به کاربران اجازه دسترسی به منابع داده‌های متمرکز شده را از روی دسک تاپ می‌دهد.

- و افزایش توانایی به تلفیق داده از منابع غیر متناجس به یک منبع قابل جستجو می‌باشد.

علاوه بر پیشرفت ابزارهای مدیریت داده، افزایش قابلیت دسترسی به داده و کاهش نرخ نگهداری داده نقش ایفا می‌کند. در طول چند سال گذشته، افزایش سریع جمع‌آوری و نگهداری حجم اطلاعات وجود داشته است. با پیشنهادها و برخی از ناظران مبنی بر آنکه کمیت داده‌های دنیا به طور تخمینی هر ساله دو برابر می‌گردد. در همین زمان هزینه ذخیره‌سازی داده‌ها بطور قابل توجهی از دلار برای هر مگابایت به پنی برای مگابایت کاهش پیدا کرده است. مطابقاً قدرت محاسبه‌ها در هر ۱۸ - ۲۴ ماه به دو برابر ارتقاء پیدا کرده است این در حالی است که هزینه قدرت محاسبه رو به کاهش است. داده‌کاوی به طور معمول در دو حوزه خصوصی و عمومی افزایش پیدا کرده است. سازمانها داده‌کاوی را به عنوان ابزاری برای بازدید اطلاعات مشتریان، کاهش تقلب و اتلاف و کمک به تحقیقات پزشکی استفاده می‌کنند. با این همه ازدیاد داده‌کاوی به طبع بعضی از پیاده‌سازی، پیامد اشتباه هم دارد. اینها شامل نگرانی‌هایی در مورد کیفیت داده‌ای که تحلیل می‌گردد، توانایی کار گروهی پایگاه‌های داده و نرم‌افزارها بین ارگانها و تخطی‌های بالقوه به حریم شخصی می‌باشد. همچنین ملاحظات در مورد محدودیت‌هایی در داده‌کاوی در ارگانها که کارشان تاثیر بر امنیت دارد، نادیده گرفته می‌شود.

یک برنامه کاربردی داده‌کاوی معمولاً با تشخیص حوزه مورد اعمال توسط داده‌کاو^۱ شروع می‌گردد. داده‌کاو، جزء نرم‌افزاری است که منابع اطلاعاتی و داده مناسب را تشخیص می‌دهد. با این داده‌های مناسب، داده‌کاوی قابل انجام است. داده‌کاوی در سه مرحله صورت می‌پذیرد:

– پیش‌پردازش^۱

داده‌های خام بنا به دلایل متفاوت برای داده‌کاوی مناسب نیستند و بایستی به منظور حذف شلوغی و بی‌قاعدگی‌ها پاک شوند. داده‌ها ممکن است حجیم و یا شامل خصیصه‌های نامربوطی باشند که در این صورت با تنزل و نمونه‌برداری از آنها، برای داده‌کاوی آماده می‌گردند.

– داده‌کاوی

در این مرحله، داده‌های پردازش شده مرحله قبل به الگوریتم‌های داده‌کاوی تغذیه می‌شوند و نهایتاً الگوها و دانش مناسب تولید می‌گردد.

– پس‌پردازش^۲

در بسیاری از کاربردها همه الگوهای کشف شده مفید نیستند در این مرحله الگوهای سودمند برای کاربردهای خاص تشخیص داده می‌شوند. ارزیابی‌های متنوع و تکنیک‌های بصری‌سازی به جهت تصمیم‌گیری استفاده می‌گردند. تمام مراحل فوق که تقریباً همیشه به صورت تکراری می‌باشند فرآیند داده‌کاوی نام دارد. در داده‌کاوی سستی از داده‌های ساخت‌یافته ذخیره شده در جدول ارتباطی، صفحه گسترده‌ها، فایل‌های یکنواخت به فرم جدولی استفاده می‌شد با رشد وب و مستندات متنی، وب‌کاوی^۳ و متن‌کاوی^۴ به‌طور فزاینده‌ای مهم و محبوب گردیدند.

۱-۲ محدودیت‌های داده‌کاوی

محصولات داده‌کاوی ابزارهای قدرتمندی می‌باشند، اما در نوع کاربردی کافی نیستند. برای کسب موفقیت، داده‌کاوی نیازمند تحلیل گران حرفه‌ای و متخصصان ماهری می‌باشد که بتوانند

-
1. pre-processing
 2. post-processing
 3. Web Mining
 4. Text Mining

ترکیب خروجی بوجود آمده را تحلیل و تفسیر نمایند. در نتیجه محدودیت‌های داده‌کاوی مربوط به داده اولیه یا افراد است تا اینکه مربوط به تکنولوژی باشد.

اگرچه داده‌کاوی به الگوهای مشخص و روابط آنها کمک می‌کند، اما برای کاربر اهمیت و ارزش این الگوها را بیان نمی‌کند. تصمیماتی از این قبیل بر عهده خود کاربر است. برای نمونه در ارزیابی صحت داده‌کاوی، برنامه کاربردی در تشخیص مظنونان تروریست طراحی شده که ممکن است این مدل به کمک اطلاعات موجود در مورد تروریست‌های شناخته شده، آزمایش شود. با این همه در حالی که ممکن است اطلاعات شخص بطور معین دوباره تصدیق گردد، که این مورد به این منظر نیست که برنامه مظنونی را که رفتارش به طور خاص از مدل اصلی منحرف شده را تشخیص بدهد.

تشخیص رابطه بین رفتارها و یا متغیرها یکی دیگر از محدودیت‌های داده‌کاوی می‌باشد که لزوماً روابط اتفاقی را تشخیص نمی‌دهد. برای مثال برنامه‌های کاربردی ممکن است الگوهای رفتاری را مشخص کند، مثل تمایل به خرید بلیط هواپیما درست قبل از حرکت که این موضوع به مشخصات درآمد، سطح تحصیلی و استفاده از اینترنت بستگی دارد. در حقیقت رفتارهای شخصی شامل شغل (نیاز به سفر در زمانی محدود) وضع خانوادگی (نیاز به مراقبت پزشکی برای مریض) یا تفریح (سود بردن از تخفیف دقایق پایانی برای دیدن مکان‌های جدید) ممکن است بر روی متغیرهای اضافه تاثیر بگذارد.

۱-۳ خلاصه فصل‌ها

این کتاب موضوعات اصلی داده‌کاوی را پوشش می‌دهد که در ادامه معرفی مفاد هر فصل آمده است و کتاب دیگری تحت عنوان وب‌کاوی موضوعات ساختارکاوی، محتواکاوی و کاربردکاوی را پوشش می‌دهد. در این کتاب وابستگی فصول محدود است به جزء فصل پنجم که بعضی تکنیک‌های فصل چهارم را استفاده می‌کند.

فصل ۲- قوانین انجمنی و الگوهای ترتیبی^۱

این فصل دو مدل داده‌کاوی مهم که در بسیاری از وظایف وب‌کاوی به خصوص کاربرد کاوی و محتوا کاوی استفاده می‌شوند را مورد بررسی قرار می‌دهد. کاوش قوانین انجمنی، مجموعه‌ای از آیت‌های داده‌ای که به طور متوالی با یکدیگر رخ می‌دهند را می‌یابد و الگوهای ترتیبی مجموعه‌ای از آیت‌های داده‌ای که با یکدیگر در بعضی مراحل یا دنباله‌ها رخ می‌دهند را می‌یابد. به عنوان مثال، در کاربرد کاوی وب، کاوش قوانین انجمنی در پیدا کردن سرکشی کاربران و الگوهای خرید مورد استفاده قرار می‌گیرد و الگوهای ترتیبی در پیدا کردن الگوهای هدایتی کاربران استفاده می‌گردند.

فصل ۳- یادگیری نظارت شده^۲

در این روش یک مجموعه آموزشی در نظر گرفته می‌شود و یادگیرنده بر اساس یک ورودی عمل کرده و یک خروجی به دست می‌آورد سپس این خروجی توسط یک سرپرست که می‌تواند خروجی مورد نظر ما باشد مورد ارزیابی قرار می‌گیرد و بر اساس اختلافی که با خروجی مطلوب دارد یک سری تغییرات در عملکرد یادگیرنده به وجود می‌آید. یادگیری نظارت شده اغلب در تکنیک یادگیری یا کاوش داده‌کاوی و وب‌کاوی به کار گرفته می‌شود. یادگیری نظارت شده همچنین طبقه‌بندی^۳ نیز نامیده می‌شود که هدف آن یادگیری تابع طبقه‌بندی (طبقه‌بندی‌کننده)^۴ از داده‌های برچسب‌دار با طبقه‌های از پیش تعریف شده می‌باشد. نتیجه طبقه‌بندی‌کننده در طبقه‌بندی کردن نمونه‌های داده آینده استفاده می‌گردد. به واسطه این واقعیت که نمونه‌های داده به کار رفته در یادگیری (داده‌های آموزشی)^۵ توسط کلاس‌های از پیش تعریف شده برچسب می‌خورند این روش یادگیری نظارت شده نامیده می‌شود.

-
1. Association Rules and Sequential Patterns
 2. Supervised Learning
 3. classification
 4. classifier
 5. Training data

فصل ۴- یادگیری بدون نظارت^۱

در یادگیری بدون نظارت، برای اطلاعات استفاده شده جهت یادگیری هیچ کلاس یا طبقه از پیش تعیین شده‌ای وجود ندارد. الگوریتم یادگیری بایستی ساختارها و قاعده‌های (نظم و ترتیب) مخفی در داده‌ها را بیابد یکی از تکنیک‌های یادگیری بدون نظارت، دسته‌بندی^۲ نام دارد در دسته‌بندی نمونه‌های داده بر اساس شباهت‌ها یا تفاوت‌هایشان در گروه‌ها و دسته‌ها سازماندهی می‌گردند. دسته‌بندی در بسیاری موارد وب‌کاوی استفاده می‌شود به عنوان مثال، می‌توانیم صفحات وب را بر اساس گروه‌هایی که هر گروه بیانگر موضوع ویژه‌ای است، دسته‌بندی کنیم. همچنین می‌توانیم مستندات را در دسته‌های سلسله مراتبی دسته‌بندی کنیم.

فصل ۵- یادگیری نظارتی جزئی^۳

یادگیری نظارت شده نیازمند تعداد زیادی از نمونه‌های داده برچسب خورده جهت آموزش یک دسته‌کننده دقیق می‌باشد. برچسب‌گذاری که اغلب به صورت دستی صورت می‌پذیرد، یک کار مشتاقانه، متمرکز و زمان‌بر است. به جهت کاهش تلاش برچسب‌گذاری دستی، یادگیری از نمونه‌های برچسب خورده و بدون برچسب^۴ که در آن یک مجموعه کوچک از نمونه‌های برچسب خورده و یک مجموعه بزرگ از نمونه‌های بدون برچسب برای یادگیری استفاده می‌شود، پیشنهاد می‌شود. این مدل همچنین یادگیری نیمه نظارت شده^۵ نامیده می‌شود.

مدل یادگیری دیگری که در این فصل بررسی می‌شود یادگیری از نمونه‌های واقعی و بدون برچسب^۶ نام دارد که برای طبقه‌بندی دو کلاس استفاده می‌گردد. اما هیچ نمونه برچسب‌دار منفی و غیرحقیقی برای یادگیری وجود ندارد این مدل برای موارد زیادی کاربرد دارد. به عنوان مثال، ما مجموعه‌ای از مقاله‌های وب‌کاوی داریم و قصد تشخیص دیگر مقالات وب‌کاوی در

1. Unsupervised Learning

2. Clustering

3. Partially Supervised Learning

4. learning from Labeled and Unlabeled examples (or LU learning)

5. Semi-Supervised learning

6. learning from Positive and Unlabeled examples (or PU learning)

مخزن مقاله‌های جستجو شده که شامل انواع مقالات می‌باشد را داریم. مجموعه مقالات وب‌کاوی به عنوان داده‌های واقعی و دیگر مقاله‌های مخزن مقالات به عنوان داده‌های بدون برچسب تلقی می‌گردند.

۱-۴ نحوه مطالعه کتاب و توصیه به اساتید

این کتاب به جهت ارائه تعلیمی محیط‌های آکادمیک برای یک ترم تحصیلی مناسب است و مناسب فارغ‌التحصیلان و دانشجویان سال آخر رشته‌های مهندسی کامپیوتر، علوم کامپیوتر و آمار می‌باشد. همچنین می‌تواند مورد استفاده پژوهشگران علاقمند به کار کردن در زمینه‌های وب‌کاوی و یا متن‌کاوی قرار گیرد.