داده‌کاوی وب

نویسنده:

بینگ لیو

مترجمین:

دکتر پریناز اسکندریان

دکتر امیر ضیاءصابونچی

دکتر فرهاد عبدزاده

انتشارات دانشگاه غیرانتفاعی آفاق

فهرست مطالب

۱. مقدمه

وقتی این کتاب را می‌خوانید، واضح است که درک کاملی از شبکه جهانی وب دارید و به طور گسترده از آن استفاده کرده‌اید. وب جهانی (یا به‌اختصار وب) تقریباً بر هر جنبه‌ای از زندگی روزمره ما تأثیر عمیقی داشته است. این شبکه بزرگ‌ترین و شناخته‌شده‌ترین منبع اطلاعاتی است که به‌راحتی قابل‌دسترسی و جستجو می‌باشد. شبکه وب از اسناد به‌هم‌پیوسته بی‌شماری (معروف به صفحات وب) که توسط میلیون‌ها نفر ایجاد شده‌اند، تشکیل شده است. از زمان آغاز، وب تأثیر عمیقی بر نحوه جستجوی اطلاعات ما داشته است. قبل از ظهور وب، جستجوی اطلاعات معمولاً شامل جستجوی راهنمایی از یک دوست یا متخصص، یا تهیه و خرید کتاب بود. اما امروزه به لطف وب، همه چیز با چند کلیک در دسترس ماست و می‌توانیم به‌راحتی از خانه یا دفتر کار خود اطلاعات موردنیاز را پیدا کرده و دانش خود را با دیگران به اشتراک بگذاریم.

وب همچنین به یک کانال مهم برای انجام کسب‌وکار تبدیل شده است. اکنون می‌توانیم تقریباً هر چیزی را از فروشگاه‌های آنلاین خریداری کنیم بدون اینکه نیاز به مراجعه فیزیکی به یک فروشگاه داشته باشیم. وب همچنین راهی مفید برای تعامل با یکدیگر، بیان افکار و ایده‌های خود و گفتگو با مردم در هر کجای جهان ارائه می‌دهد. وب در واقع یک جامعه مجازی است. ما وب، پیشینه آن و موضوعاتی را که در این کتاب در این فصل اول پوشش خواهیم داد، ارائه می‌کنیم.

۱.۱. شبکه جهانی وب چیست؟

شبکه جهانی وب به طور رسمی به‌عنوان "ابتکاری برای بازیابی اطلاعات چندرسانه‌ای در یک محدوده وسیع که هدف آن ارائه دسترسی جهانی به مجموعه‌ای بزرگ از اسناد است" تعریف می‌شود. به بیان ساده‌تر، وب یک شبکه رایانه‌ای مبتنی بر اینترنت است که به کاربران اجازه می‌دهد تا از طریق شبکه جهانی اینترنت به اطلاعات ذخیره‌شده در رایانه‌های دیگر دسترسی پیدا کنند.

پیاده‌سازی وب به یک مدل مرسوم مشتری-سرور پایبند است. در این مدل، کاربر از یک برنامه (که کلاینت نامیده می‌شود) برای اتصال به یک ماشین راه دور (که سرور نامیده می‌شود) که داده‌ها در آن ذخیره شده‌اند، استفاده می‌کند. پیمایش در وب از طریق یک برنامه کلاینت به نام مرورگر انجام

می‌شود، مانند نت اسکیپ، اینترنت اکسپلورر، فایرفاکس، کروم و غیره. مرورگرهای وب با ارسال درخواست به سرورهای راه دور برای دریافت اطلاعات و سپس تفسیر اسناد بازگردانده شده که به زبان HTML نوشته شده‌اند، متن و گرافیک را بر روی صفحه رایانه کاربر نمایش می‌دهند.

عملکرد وب بر پایه ساختار اسناد ابر متنی (Hypertext) استوار است. ابر متن به نویسندگان صفحات وب این امکان را می‌دهد که اسناد خود را به دیگر اسناد مرتبط که در رایانه‌های مختلف در سراسر جهان قرار دارند، پیوند دهند. برای مشاهده این اسناد، کاربر تنها کافی است پیوندها (که به آن‌ها "هایپرلینک" گفته می‌شود) را دنبال کند.

ایده ابر متن در سال ۱۹۶۵ توسط تد نلسون ابداع شد که سیستم مشهور ابر متنی زانادو[1] را نیز ایجاد کرد (http://xanadu.com/). "چندرسانه‌ای" (Hypermedia) به ابر متنی گفته می‌شود که از انواع رسانه‌های مختلف مانند فایل‌های صوتی، تصویری و تصویری پشتیبانی می‌کند.

۱.۲. تاریخچه مختصر وب و اینترنت

اختراع وب: وب در سال ۱۹۸۹ توسط تیم برنرز-لی، که در آن زمان در مرکز تحقیقات هسته‌ای اروپا (CERN) در سوئیس کار می‌کرد، اختراع شد. او واژه "شبکه جهانی وب" را ابداع کرد، اولین سرور وب به نام "httpd" و اولین برنامه کلاینت (یک مرورگر و ویرایشگر) به نام "WorldWideWeb" را نوشت. این پروژه در مارس ۱۹۸۹ آغاز شد، زمانی که تیم برنرز-لی پیشنهادی با عنوان "مدیریت اطلاعات " به مدیران خود در CERN ارائه کرد. در این پیشنهاد، او به معایب سازماندهی سلسله‌مراتبی اطلاعات اشاره کرده و مزایای یک سیستم مبتنی بر ابر متن را توضیح داد. این پیشنهاد شامل یک پروتکل ساده برای درخواست اطلاعات ذخیره‌شده در سیستم‌های کامپیوتری از طریق شبکه‌ها و یک الگوی تبادل اطلاعات در قالبی مشترک بود، به‌طوری که اسناد افراد می‌توانستند از طریق پیوندهای ابر متنی به سایر اسناد متصل شوند. همچنین روش‌هایی برای خواندن متن و گرافیک

[1] xanadu

با استفاده از فناوری نمایشگر در CERN در آن زمان پیشنهاد شد. این پیشنهاد یک سیستم فرامتن توزیع شده را توصیف می‌کند که ساختار بنیادی وب را تشکیل می‌دهد.

ابتدا، این پیشنهاد حمایت لازم را دریافت نکرد. اما در سال ۱۹۹۰، برنرز–لی دوباره پیشنهاد خود را مطرح کرد و این بار حمایت لازم برای شروع کار را به دست آورد. با این پروژه، برنرز–لی و تیم او در CERN پایه‌های توسعه آینده وب به‌عنوان یک سیستم ابر متن توزیع‌شده را بنا نهادند. آنها سرور و مرورگر خود، پروتکل ارتباطی بین کلاینت و سرور یعنی پروتکل انتقال ابر متن (HTTP)، زبان نشانه‌گذاری ابر متن (HTML) برای نگارش اسناد وب، و آدرس‌های یکتای منبع (URL) را معرفی کردند. و این‌گونه بود که وب آغاز شد.

ظهور مرورگرهای موزاییک[1] و نت اسکیپ: یکی از رویدادهای مهم در توسعه وب، ظهور مرورگر موزاییک بود. در فوریه ۱۹۹۳، مارک اندریسن از مرکز ملی برنامه‌های کاربردی ابررایانه (NCSA) در دانشگاه ایلینوی و تیمش اولین مرورگر گرافیکی وب برای سیستم‌عامل یونیکس با نام "Mosaic for X" را عرضه کردند. چند ماه بعد، نسخه‌های دیگری از موزاییک برای سیستم‌های عامل مکینتاش و ویندوز نیز منتشر شد. این رویداد به دلیل ایجاد یک رابط کاربری گرافیکی ساده و یکپارچه برای سه سیستم‌عامل محبوب آن زمان، بسیار مهم بود. این مرورگر به‌سرعت از دایره دانشگاهی فراتر رفت و در میانه دهه ۹۰ میلادی، جیم کلارک، بنیان‌گذار شرکت سیلیکون گرافیکس، با همکاری مارک اندریسن شرکت موزاییک کامیونیکیشنز (که بعدها به نت اسکیپ کامیونیکیشنز تغییر نام داد) را تأسیس کردند. در عرض چند ماه، مرورگر نت اسکیپ به عموم عرضه شد و رشد سریع وب آغاز شد. مرورگر اینترنت اکسپلورر از مایکروسافت نیز در آگوست ۱۹۹۵ وارد بازار شد و به رقابت با نت اسکیپ پرداخت. ایجاد شبکه جهانی وب توسط تیم برنرز لی و پس از آن عرضه مرورگر موزاییک، به‌عنوان دو عامل اصلی در موفقیت و محبوبیت وب شناخته می‌شوند.

اینترنت: وب بدون وجود اینترنت امکان‌پذیر نبود، چرا که اینترنت شبکه ارتباطی لازم برای عملکرد وب را فراهم می‌کند. اینترنت با شبکه کامپیوتری آرپانت در دوران جنگ سرد آغاز شد. این پروژه در

[1] Mosaic

ایالات متحده باهدف حفظ کنترل بر موشک‌ها و بمب‌افکن‌ها پس از یک حمله هسته‌ای به وجود آمد و توسط آژانس پروژه‌های تحقیقاتی پیشرفته (ARPA)، که بخشی از وزارت دفاع ایالات متحده بود، حمایت شد. اولین ارتباطات آرپانت در سال ۱۹۶۹ برقرار شد و در سال ۱۹۷۲، در اولین کنفرانس بین‌المللی کامپیوتر و ارتباطات در واشنگتن‌دی‌سی به نمایش گذاشته شد. در این کنفرانس، دانشمندان آرپا کامپیوترهایی از ۴۰ نقطه مختلف را به هم متصل کردند. در سال ۱۹۷۳، وینتون سرف و باب کان توسعه پروتکلی را آغاز کردند که بعدها به نام پروتکل TCP/IP شناخته شد. سال بعد، آن‌ها مقاله‌ای با عنوان "پروتکل کنترل انتقال" منتشر کردند که آغازگر TCP/IP بود. این پروتکل جدید به شبکه‌های کامپیوتری مختلف امکان می‌داد تا به یکدیگر متصل و با هم ارتباط برقرار کنند. در سال‌های بعد، شبکه‌های متعددی ایجاد شد و تکنیک‌ها و پروتکل‌های مختلفی توسعه یافت. بااین‌حال، آرپانت همچنان به‌عنوان ستون فقرات سیستم باقی ماند. در آن دوران، فضای شبکه‌ها بسیار پراکنده و آشفته بود. نهایتاً در سال ۱۹۸۲، پروتکل TCP/IP به طور رسمی پذیرفته شد و اینترنت، که مجموعه‌ای متصل از شبکه‌ها با استفاده از پروتکل TCP/IP بود، به وجود آمد.

موتور جستجو: با گسترش روزافزون اطلاعات در سطح جهان، نیاز به سیستمی برای جستجو و دسترسی منظم و کارآمد به این اطلاعات بیش‌ازپیش احساس شد. این نیاز به توسعه موتورهای جستجو منجر شد. در سال ۱۹۹۳، گروهی از دانشجویان دانشگاه استنفورد سیستم جستجوی Excite را معرفی کردند. در سال ۱۹۹۴، موتور جستجوی EINet Galaxy به‌عنوان بخشی از پروژه تحقیقاتی MCC در دانشگاه تگزاس ایجاد شد. در همان سال، جری یانگ و دیوید فیلو یاهو! را تأسیس کردند که در ابتدا فهرستی از وب‌سایت‌های موردعلاقه آن‌ها بود و جستجوی دایرکتوری ارائه می‌کرد. طی سال‌های بعد، موتورهای جستجوی دیگری همچون Lycos، Inforseek، AltaVista، Inktomi و Ask Jeeves وارد عرصه شدند. گوگل نیز در سال ۱۹۹۸ توسط سرگئی برین و لری پیج، بر اساس پروژه تحقیقاتی آن‌ها در دانشگاه استنفورد، راه‌اندازی شد. مایکروسافت نیز در سال ۲۰۰۳ وارد حوزه جستجو شد و در بهار ۲۰۰۵ موتور جستجوی MSN (که اکنون Bing نامیده می‌شود) را عرضه کرد. یاهو! در سال ۲۰۰۴ پس از خرید Inktomi در سال ۲۰۰۳، قابلیت جستجوی عمومی را به سرویس خود افزود.

کنسرسیوم جهانی وب (W3C) : در دسامبر ۱۹۹۴، کنسرسیوم جهانی وب (W3C) توسط MIT و CERN به‌عنوان یک سازمان بین‌المللی برای هدایت توسعه وب تشکیل شد. هدف اصلی W3C "ترویج استانداردها برای تکامل وب و ایجاد قابلیت همکاری بین محصولات WWW از طریق تولید مشخصات و نرم‌افزارهای مرجع" بود. اولین کنفرانس بین‌المللی وب جهانی (WWW) نیز در همان سال برگزار شد و از آن زمان به طور سالانه ادامه یافته است. از سال ۱۹۹۵ تا ۲۰۰۱، وب به‌سرعت رشد کرد و سرمایه‌گذاران با دیدن فرصت‌های تجاری، به آن علاقه‌مند شدند. کسب‌وکارهای متعددی در وب شروع به فعالیت کردند که منجر به توسعه‌های غیرمنطقی شد. این حال، توسعه وب دیگر متوقف نشد و تنها به شکلی منطقی‌تر ادامه یافت.

۱.۳ داده‌کاوی وب

در دهه گذشته، وب رشد قابل‌توجهی را تجربه کرده است و خود را به‌عنوان بزرگ‌ترین منبع داده در دسترس عموم در سطح جهان معرفی کرده است. شبکه وب دارای ویژگی‌های متمایز متعددی است که فرایند استخراج اطلاعات و دانش باارزش را جذاب و چالش‌برانگیز می‌کند. اکنون می‌توانیم نگاهی دقیق‌تر به برخی از این ویژگی‌ها بیندازیم.

۱. *حجم داده‌ها و اطلاعات موجود در وب بسیار زیاد است و همچنان در حال گسترش است. گستره اطلاعات نیز بسیار وسیع و متنوع است؛ به‌طوری‌که می‌توان درباره هر موضوعی در وب اطلاعاتی یافت.*
۲. *طیف گسترده‌ای از داده‌ها در وب در دسترس است، از جمله جداول ساختاریافته، صفحات نیمه ساختاریافته، متون بدون ساختار، و فایل‌های چندرسانه‌ای مانند تصاویر، فایل‌های صوتی و ویدئو.*
۳. *اطلاعات موجود در وب به دلیل نویسندگان متنوع و قالب‌های مختلف صفحات، بسیار ناهمگون است. بسیاری از صفحات وب ممکن است اطلاعات مشابهی را با استفاده از کلمات و فرمت‌های کاملاً متفاوت ارائه دهند که این امر باعث می‌شود یکپارچه‌سازی اطلاعات از چندین صفحه به یک مشکل چالش‌برانگیز تبدیل شود.*
۴. *همچنین، حجم قابل‌توجهی از اطلاعات در وب به هم مرتبط هستند. لینک‌ها بین صفحات وب درون یک سایت و بین سایت‌های مختلف وجود دارند. درون یک سایت، این لینک‌ها به‌عنوان یک مکانیزم سازماندهی اطلاعات عمل می‌کنند و بین سایت‌های مختلف، این لینک‌ها به طور*

ضمنی به صفحات هدف اعتبار می‌بخشند. صفحاتی که توسط تعداد زیادی از صفحات دیگر لینک می‌شوند، معمولاً به‌عنوان صفحات باکیفیت یا معتبر شناخته می‌شوند، زیرا بسیاری از افراد به آن‌ها اعتماد دارند.

۵. *اطلاعات موجود در وب اغلب پر از نویز است. این نویز از دو منبع اصلی ناشی می‌شود: اول اینکه یک صفحه وب معمولی شامل چندین بخش اطلاعاتی مختلف است، مانند محتوای اصلی، لینک‌های ناوبری، تبلیغات، اخطارهای حق‌تألیف و سیاست‌های حفظ حریم خصوصی. برای یک کاربرد خاص، تنها بخشی از این اطلاعات مفید است و بقیه به‌عنوان نویز تلقی می‌شوند. برای تحلیل دقیق اطلاعات وب و استخراج داده‌ها، لازم است این نویز حذف شود. دوم اینکه به دلیل نبود کنترل کیفی در وب، هر فرد می‌تواند هرچه که می‌خواهد را منتشر کند که منجر به حجم زیادی از اطلاعات بی‌کیفیت، نادرست یا حتی گمراه‌کننده می‌شود.*

۶. *علاوه بر این، وب بستری برای کسب‌وکار و تجارت است. تمامی وب‌سایت‌های تجاری به افراد این امکان را می‌دهند که عملیات مفیدی را انجام دهند، مانند خرید محصولات، پرداخت قبوض و پر کردن فرم‌ها. برای پشتیبانی از چنین کاربردهایی، وب‌سایت‌ها باید انواع مختلفی از خدمات خودکار، مانند سیستم‌های پیشنهاددهی، ارائه دهند.*

۷. *وب همچنین پویاست و اطلاعات موجود در آن به طور مداوم در حال تغییر است. به‌روزرسانی و نظارت بر این تغییرات، برای بسیاری از کاربردها اهمیت دارد.*

۸. *در نهایت، وب یک جامعه مجازی است. وب تنها شامل داده‌ها، اطلاعات و خدمات نیست، بلکه درباره تعاملات بین افراد، سازمان‌ها و سیستم‌های خودکار نیز هست. افراد می‌توانند به‌راحتی و بلافاصله با دیگران در سراسر جهان ارتباط برقرار کنند و نظرات و دیدگاه‌های خود را در فروم‌های اینترنتی، وبلاگ‌ها، سایت‌های بررسی و شبکه‌های اجتماعی بیان کنند. این اطلاعات جدید امکان انجام بسیاری از وظایف داده‌کاوی جدید مانند تحلیل نظرات و تحلیل شبکه‌های اجتماعی را فراهم می‌کنند.*

تمامی این ویژگی‌ها، چالش‌ها و فرصت‌هایی را برای استخراج و کشف اطلاعات و دانش از وب ایجاد می‌کنند. در این کتاب، ما تنها بر روی استخراج داده‌های متنی تمرکز می‌کنیم. برای استخراج تصاویر، ویدئوها و صداها به منابع [۱۵, ۲۶] مراجعه کنید. برای بررسی استخراج اطلاعات از وب، لازم است با داده‌کاوی آشنا شویم، چرا که این تکنیک در بسیاری از وظایف مرتبط با استخراج از وب مورداستفاده قرار گرفته است. بااین‌حال، استخراج اطلاعات از وب صرفاً به‌کارگیری داده‌کاوی نیست. به دلیل غنای

اطلاعات و تنوع آن‌ها و دیگر ویژگی‌های خاص وب که پیش‌تر به آن‌ها اشاره شد، استخراج از وب بسیاری از الگوریتم‌های خاص خود را توسعه داده است.

۱.۳.۱ داده‌کاوی چیست؟

داده‌کاوی یا به عبارتی استخراج داده‌ها که به آن کشف دانش در پایگاه‌های داده (KDD) نیز گفته می‌شود، به‌عنوان فرایند کشف الگوها یادانش مفید از منابع مختلف داده؛ مانند پایگاه‌های داده، متون، تصاویر و وب شناخته می‌شود. الگوهایی که استخراج می‌شوند باید معتبر، بالقوه مفید و قابل‌درک باشند. استخراج داده‌ها یک حوزه چندرشته‌ای است که شامل یادگیری ماشین، آمار، پایگاه‌های داده، هوش مصنوعی، بازیابی اطلاعات و تجسم داده‌ها می‌شود.

وظایف مختلفی در استخراج داده‌ها وجود دارد، که برخی از رایج‌ترین آنها شامل یادگیری نظارت شده (یا طبقه‌بندی)، یادگیری بدون نظارت (یا خوشه‌بندی)، استخراج قوانین انجمنی و استخراج الگوهای متوالی است. در این کتاب به بررسی همه این وظایف خواهیم پرداخت.

کاربرد استخراج داده‌ها معمولاً با فهم دامنه کاربرد توسط تحلیل‌گران داده (استخراج‌کنندگان داده) آغاز می‌شود، سپس منابع داده مناسب و داده‌های هدف شناسایی می‌شوند. پس از جمع‌آوری داده‌ها، فرایند استخراج داده‌ها در سه مرحله اصلی انجام می‌شود:

– **پیش‌پردازش:** داده‌های خام معمولاً به دلیل وجود نویز یا ناهنجاری‌ها برای استخراج مناسب نیستند و باید پاک‌سازی شوند. همچنین ممکن است داده‌ها بسیار بزرگ باشند یا شامل ویژگی‌های غیرمرتبط زیادی باشند که نیاز به کاهش داده از طریق نمونه‌گیری و انتخاب ویژگی دارند.

– **استخراج داده:** داده‌های پردازش‌شده به یک الگوریتم استخراج داده، داده می‌شود که الگوها یادانش موردنظر را تولید می‌کند.

– پس‌پردازش: در بسیاری از کاربردها، همه الگوهای کشف‌شده مفید نیستند. این مرحله به شناسایی الگوهای مفید برای کاربردهای مختلف می‌پردازد. تکنیک‌های مختلف ارزیابی و تجسم برای این منظور به کار گرفته می‌شوند.

این فرایند که به آن فرایند استخراج داده نیز گفته می‌شود، معمولاً به‌صورت تکراری انجام می‌شود و ممکن است چندین بار تکرار شود تا به نتیجه نهایی و رضایت‌بخش دست یابد که سپس در وظایف عملیاتی دنیای واقعی مورداستفاده قرار می‌گیرد.

در استخراج داده‌های سنتی، از داده‌های ساختاریافته‌ای استفاده می‌شود که در جداول رابطه‌ای، صفحات گسترده یا فایل‌های تخت به شکل جدول ذخیره شده‌اند. با رشد وب و اسناد متنی، استخراج داده از وب و استخراج متون اهمیت و محبوبیت بیشتری پیدا کرده‌اند. تمرکز این کتاب بر استخراج داده از وب است.

۱.۳.۲ وب ماینینگ چیست؟

وب ماینینگ(**وب‌کاوی**) یا استخراج داده از وب به‌منظور کشف اطلاعات مفید یادانش از ساختار لینک‌ها، محتوای صفحات وب و داده‌های مربوط به استفاده کاربران انجام می‌شود. اگرچه وب ماینینگ از تکنیک‌های مختلف داده‌کاوی بهره می‌برد، اما به دلیل ناهمگونی و طبیعت نیمه‌ساختاریافته یا غیرساختاریافته داده‌های وب، نمی‌توان آن را به طور کامل به‌عنوان یک کاربرد ساده از تکنیک‌های سنتی داده‌کاوی در نظر گرفت. در دهه اخیر، وظایف و الگوریتم‌های جدیدی برای وب ماینینگ توسعه یافته‌اند. این وظایف بر اساس نوع داده‌هایی که در فرایند استخراج مورداستفاده قرار می‌گیرند، به سه دسته اصلی تقسیم می‌شوند: استخراج ساختار وب، استخراج محتوا از وب و استخراج اطلاعات از رفتار کاربران وب.

۱. استخراج ساختار وب: این نوع استخراج به کشف دانش مفید از هایپرلینک‌ها (لینک‌ها) که ساختار وب را تشکیل می‌دهند، می‌پردازد. به‌عنوان‌مثال، از طریق لینک‌ها می‌توان صفحات وب مهم را شناسایی کرد، که این فناوری در موتورهای جستجو مورداستفاده قرار می‌گیرد. همچنین می‌توان جوامع کاربری با علاقه‌مندی‌های مشترک را کشف کرد. این وظایف در داده‌کاوی سنتی انجام نمی‌شوند، زیرا معمولاً ساختار لینک‌ها در جداول رابطه‌ای وجود ندارد.

۲. استخراج محتوا از وب: این نوع استخراج به استخراج یا استخراج اطلاعات و دانش مفید از محتوای صفحات وب می‌پردازد. به‌عنوان‌مثال، می‌توان صفحات وب را به طور خودکار بر اساس

موضوعات آن‌ها طبقه‌بندی و خوشه‌بندی کرد. این وظایف مشابه وظایف داده‌کاوی سنتی هستند. اما همچنین می‌توان الگوهایی در صفحات وب کشف کرد تا داده‌های مفیدی مانند توضیحات محصولات یا پست‌های انجمن‌ها را استخراج کرد. علاوه بر این، می‌توان نظرات مشتریان و پست‌های انجمن‌ها را بررسی کرد تا نظرات مصرف‌کنندگان را کشف کرد، که این‌ها وظایف مرسوم داده‌کاوی نیستند.

۳. استخراج اطلاعات از رفتار کاربران وب: این نوع استخراج به کشف الگوهای دسترسی کاربران از طریق لاگ‌های استفاده از وب که هر کلیک کاربر را ثبت می‌کنند، می‌پردازد. استخراج رفتار کاربران وب از بسیاری از الگوریتم‌های داده‌کاوی بهره می‌برد. یکی از مسائل کلیدی در این زمینه، پیش‌پردازش داده‌های کلیک‌استریم در لاگ‌های استفاده است تا داده‌های مناسبی برای استخراج آماده شود.

در این کتاب، ما به بررسی هر سه نوع وب ماینینگ خواهیم پرداخت. بااین‌حال، به دلیل ازدحام اطلاعات موجود در وب، تعداد زیادی از وظایف وب ماینینگ وجود دارد که نمی‌توانیم همه آن‌ها را پوشش دهیم و بر روی برخی از وظایف مهم و الگوریتم‌های اساسی آن‌ها تمرکز خواهیم کرد. فرایند وب ماینینگ مشابه فرایند داده‌کاوی است، اما تفاوت عمده آنها در جمع‌آوری داده‌ها است. در داده‌کاوی سنتی، داده‌ها معمولاً از پیش جمع‌آوری شده و در یک انبار داده ذخیره می‌شوند. اما در وب ماینینگ، جمع‌آوری داده‌ها می‌تواند یک وظیفه مهم باشد، به‌ویژه برای استخراج ساختار و محتوا از وب که نیاز به جستجوی تعداد زیادی از صفحات وب هدف دارد. ما به این موضوع یک فصل کامل اختصاص خواهیم داد.

پس از جمع‌آوری داده‌ها، فرایند سه‌مرحله‌ای مشابهی را دنبال می‌کنیم: پیش‌پردازش داده‌ها، داده‌کاوی وب و پس‌پردازش. بااین‌حال، روش‌هایی که در هر یک از این مراحل به کار گرفته می‌شوند، ممکن است تفاوت قابل‌توجهی با تکنیک‌های مورداستفاده در داده‌کاوی سنتی داشته باشند.

۱.۴ خلاصه فصل ها

این کتاب به دو بخش اصلی تقسیم می‌شود. بخش اول، شامل فصل‌های ۲ تا ۵، به مباحث کلیدی داده‌کاوی می‌پردازد. بخش دوم که شامل بقیه فصول است، به وب‌کاوی اختصاص دارد، که یک فصل

آن به جستجوی وب مربوط می‌شود. ازآنجایی‌که مرز دقیقی بین جستجوی وب و وب‌کاوی وجود ندارد، این دو موضوع در کنار هم بررسی شده‌اند. فصل‌های ۷ و ۸ به تحلیل ساختار وب می‌پردازند که ارتباط نزدیکی با جستجوی وب دارند (فصل ۶). فصل‌های ۹ تا ۱۱ به ماینینگ محتوای وب اختصاص دارند و فصل ۱۲ به ماینینگ استفاده از وب می‌پردازد.

فصل ۲ به قوانین وابستگی و الگوهای ترتیبی می‌پردازد. این فصل به بررسی دو مدل مهم داده‌کاوی می‌پردازد که در بسیاری از وظایف وب‌کاوی، به‌ویژه در ماینینگ استفاده و محتوای وب کاربرد دارند. قوانین وابستگی مجموعه داده‌هایی را که به طور مکرر با هم ظاهر می‌شوند، شناسایی می‌کند و الگوهای ترتیبی مجموعه داده‌هایی را که در یک ترتیب خاص به طور مکرر ظاهر می‌شوند، کشف می‌کند.

فصل ۳ به یادگیری نظارت شده اختصاص دارد که یکی از پرکاربردترین تکنیک‌های یادگیری در داده‌کاوی و وب‌کاوی است. در این روش، از داده‌هایی با کلاس‌های از پیش تعریف‌شده برای یادگیری یک تابع طبقه‌بندی استفاده می‌شود که می‌تواند برای طبقه‌بندی داده‌های جدید مورداستفاده قرار گیرد.

فصل ۴ به یادگیری بدون نظارت می‌پردازد، که در آن داده‌ها بدون برچسب‌های از پیش تعریف‌شده هستند و الگوریتم یادگیری باید ساختارهای پنهان در داده‌ها را کشف کند. یکی از تکنیک‌های کلیدی در یادگیری بدون نظارت، خوشه‌بندی است که داده‌ها را بر اساس شباهت‌ها یا تفاوت‌ها به گروه‌هایی تقسیم می‌کند.

فصل ۵ یادگیری نیمه نظارت شده را معرفی می‌کند که ترکیبی از یادگیری نظارت شده و بدون نظارت است. این روش به‌منظور کاهش نیاز به داده‌های برچسب‌دار، از تعداد کمی داده‌های برچسب‌دار و تعداد زیادی داده‌های بدون برچسب استفاده می‌کند.

فصل ۶ به بازیابی اطلاعات و جستجوی وب اختصاص دارد. جستجو یکی از بزرگ‌ترین کاربردهای وب است و ریشه در زمینه بازیابی اطلاعات دارد که به کاربران کمک می‌کند اطلاعات موردنیاز خود را از مجموعه‌های بزرگ متنی پیدا کنند. این فصل به بررسی چالش‌های خاص جستجوی وب می‌پردازد.

فصل ۷ تحلیل شبکه‌های اجتماعی را بررسی می‌کند. هایپرلینک‌ها، که وب‌سایت‌ها را به هم متصل می‌کنند، به‌عنوان یک ویژگی خاص در تحلیل شبکه‌های اجتماعی مورداستفاده قرار گرفته‌اند. این فصل به بررسی الگوریتم‌های تحلیل هایپرلینک مانند PageRank و HITS و همچنین الگوریتم‌های کشف جامعه می‌پردازد.

فصل ۸ به خزیدن وب اختصاص دارد. خزنده‌های وب برنامه‌هایی هستند که به‌صورت خودکار ساختار هایپرلینک‌های وب را طی کرده و صفحات وب را به‌صورت محلی ذخیره می‌کنند. این فصل به بررسی چالش‌ها و تکنیک‌های مختلف در خزیدن وب می‌پردازد.

فصل ۹ استخراج داده‌های ساختاریافته را در برمی‌گیرد. بسیاری از صفحات وب دارای داده‌های ساختاریافته‌ای هستند که معمولاً از پایگاه‌های داده استخراج شده و در صفحات وب نمایش داده می‌شوند. استخراج این داده‌ها امکان ارائه خدمات ارزش‌افزوده مانند خرید مقایسه‌ای و جستجوی فراگیر را فراهم می‌کند.

فصل ۱۰ به یکپارچه‌سازی اطلاعات می‌پردازد. برای استفاده از داده‌ها یا اطلاعات استخراج‌شده از چندین سایت به‌منظور ارائه خدمات ارزش‌افزوده، باید این داده‌ها/اطلاعات را به‌صورت معنایی یکپارچه کنیم تا یک پایگاه‌داده منسجم و یکپارچه ایجاد شود.

فصل ۱۱ ماینینگ نظرات و تحلیل احساسات مورد بررسی قرار می‌دهد. علاوه بر داده‌های ساختاریافته، وب شامل حجم زیادی از متن‌های غیرساختاریافته نیز می‌باشد. این فصل بر ماینینگ نظرات و احساسات مردم در بررسی‌های محصولات، بحث‌های انجمنی و وبلاگ‌ها تمرکز دارد.

فصل ۱۲ نحوه استفاده وب‌کاوی را توضیح می‌دهد. همچنین به مطالعه کلیک‌های کاربران و کاربردهای آن‌ها در تجارت الکترونیک و هوش تجاری می‌پردازد که هدف آن مدل‌سازی الگوهای رفتاری و پروفایل‌های کاربران است تا بتوان سازماندهی و ساختار سایت را بهبود بخشید و تجربیات شخصی‌سازی‌شده برای کاربران ایجاد کرد.

۱.۵ نحوه مطالعه کتاب