

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

داده کاوی وب

نویسنده:

بینگ لیو

مترجمین:

دکتر پریناز اسکندریان

دکتر امیر ضیاء صابونچی

دکتر فرهاد عبدزاده



سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

فهرست مطالب

۱. مقدمه	۱
۱.۱ شبکه جهانی وب چیست؟	۱
۱.۲ تاریخچه مختصر وب و اینترنت	۲
۱.۳ داده کاوی وب	۵
۱.۳.۱ داده کاوی چیست؟	۷
۱.۳.۲ وب ماینینگ چیست؟	۸
۱.۴ خلاصه فصل ها	۹
۱.۵ نحوه مطالعه کتاب	۱۱
یادداشت ها	۱۳
بخش اول: مبانی داده کاوی	۱۵
۲. قانون وابستگی و الگوهای ترتیبی	۱۶
۲.۱ مفاهیم اصلی قانون وابستگی	۱۷
۲.۲ Apriori الگوریتم	۲۰
۲.۲.۱ تولید مجموعه اقلام تکرار شونده	۲۱
۲.۲.۲ تولید قانون وابستگی	۲۶
۲.۳ فرمت های داده برای ماینینگ قانون وابستگی	۲۹
۲.۴ ماینینگ با پشتیبانی های چند گانه	۳۱
۲.۴.۱ مدل توسعه یافته	۳۴

۳۶	۲.۴.۲. الگوریتم ماینینگ
۴۳	۲.۴.۳. تولید قانون
۴۵	۲.۵. ماینینگ قانون وابستگی کلاس
۴۶	۲.۵.۱. تعریف مسئله
۴۸	۲.۵.۲. الگوریتم ماینینگ
۵۳	۲.۵.۳. ماینینگ با پشتیبانی های کمینه چند گانه
۵۴	۲.۶. مفاهیم پایه ای الگوهای ترتیبی
۵۷	۲.۷. الگوریتم GPS :
۵۸	۲.۷.۱. ماینینگ با پشتیبانی های کمینه چند گانه
۶۶	۲.۸. ماینینگ الگوهای ترتیبی بر اساس توسعه پیشوندی
۶۶	۲.۸.۱. الگوریتم توسعه پیشوندی
۶۹	۲.۸.۲. ماینینگ با پشتیبانی های کمینه چند گانه
۷۲	۲.۹. تولید قانون از الگوهای ترتیبی
۷۲	۲.۹.۱. قانون ترتیبی
۷۳	۲.۹.۲. قانون ترتیبی بر چسب
۷۵	۲.۹.۳. قانون ترتیبی کلاس
۷۷	فصل سه یادگیری نظارتی
۷۸	۳.۱. مفاهیم پایه
۸۴	۳.۲. استقرار درخت تصمیم

۸۸.....	۳.۲.۱. الگوریتم یادگیری
۸۹.....	۳.۲.۲. تابع ناخالصی
۹۵.....	۳.۲.۳. مدیریت ویژگی‌های پیوسته
۹۷.....	۳.۲.۴. برخی مسائل دیگر
۱۰۱.....	۳.۳. ارزیابی کلاس بند
۱۰۲.....	۳.۳.۱. روش‌های ارزیابی
۱۰۴.....	۳.۳.۲. صحت، فراخوانی، F-امتیاز و نقطه Breakeven
۱۰۸.....	۳.۴. استقرار قانون
۱۰۸.....	۳.۴.۱. پوشش ترتیبی
۱۱۱.....	۳.۴.۲. یادگیری قانون: تابع Learn-One-Rule
۱۱۶.....	۳.۵. کلاس‌بندی بر اساس وابستگی‌ها
۱۱۷.....	۳.۵.۱. کلاس‌بندی با استفاده از قانون وابستگی کلاس
۱۲۳.....	۳.۵.۲. قانون وابستگی کلاس به عنوان ویژگی
۱۲۳.....	۳.۵.۳. کلاس‌بندی با استفاده از قانون وابستگی معمولی
۱۲۵.....	۳.۶. کلاس‌بندی بیزین ناوی
۱۳۰.....	۳.۷. کلاس‌بندی متن بیزین ناوی
۱۳۱.....	۳.۷.۱. چهارچوب احتمالی
۱۳۳.....	۳.۷.۲. مدل بیزین ناوی
۱۳۶.....	۳.۷.۳. بحث و گفتگو

۳۸.	ماشین‌های بردار پشتیبانی	۱۳۷
۳۸.۱	SVM خطی : حالت قابل جداسازی (جداشدنی)	۱۴۰
۳۸.۲	SVM خطی : حالت غیر قابل جداسازی	۱۴۸
۳۸.۳	SVM غیرخطی : توابع هسته	۱۵۳
۳۹.	یادگیری نزدیک‌ترین همسایه	۱۵۹
۳.۱۰	جمع‌بندی کلاس‌بندها	۱۶۱
۳.۱۰.۱	bagging	۱۶۱
۳.۱۰.۲	boosting	۱۶۲
	فصل چهارم	۱۶۵
	یادگیری غیر نظارتی	۱۶۵
۴.۱	مفاهیم پایه	۱۶۷
۴.۲	خوشه‌بندی k-means	۱۷۰
۴.۲.۱	الگوریتم k-means	۱۷۱
۴.۲.۲	نسخه دیسک از الگوریتم k-means	۱۷۴
۴.۲.۳	نقاط قوت و نقاط ضعف	۱۷۵
۴.۳	نمایش خوشه‌ها	۱۸۲
۴.۳.۱	روش‌های متداول برای نمایش خوشه‌ها	۱۸۲
۴.۳.۲	خوشه‌های با اندازه دلخواه	۱۸۴
۴.۴	خوشه‌بندی سلسله‌مراتبی	۱۸۵

۱۸۸.....	۴.۴.۱. روش تک- پیوند
۱۸۹.....	۴.۴.۲. روش پیوند کامل
۱۹۰.....	۴.۴.۳. روش پیوند میانگین
۱۹۱.....	۴.۴.۴. نقاط ضعف و قوت
۱۹۲.....	۴.۵. تابع فاصله
۱۹۲.....	۴.۵.۱. ویژگی‌های عددی
۱۹۳.....	۴.۵.۲. ویژگی‌های اسمی و باینری
۱۹۶.....	۴.۵.۳. مستندات متنی
۱۹۶.....	۴.۶. استانداردسازی داده‌ها
۲۰۰.....	۴.۷. مدیریت ویژگی‌های ترکیبی
۲۰۲.....	۴.۸. از کدام الگوریتم خوشه‌بندی باید استفاده کرد؟
۲۰۳.....	۴.۹. ارزیابی خوشه
۲۰۷.....	۴.۱۰. بازیابی نواحی داده و چاله‌ها
۲۱۱.....	فصل پنجم
۲۱۱.....	یادگیری نظارتی جزئی
۲۱۳.....	۵.۱. یادگیری از روی نمونه‌های برچسب‌گذاری شده و برچسب‌گذاری نشده
۲۱۴.....	۵.۱.۱. الگوریتم EM یا کلاس‌بندی بیزین ناوی
۲۱۹.....	۵.۱.۲. co- training
۲۲۳.....	۵.۱.۳. Self-Training

۲۲۳.....	۵.۱.۴. ماشین‌های بردار پشتیبانی Transductive
۲۲۶.....	۵.۱.۵. روش‌های مبتنی بر گراف
۲۳۱.....	۵.۱.۶. بحث و گفتگو
۲۳۲.....	۵.۲. یادگیری از روی نمونه‌های مثبت و برچسب‌گذاری نشده
۲۳۳.....	۵.۲.۱. کاربردهای یادگیری PU
۲۳۴.....	یادگیری با مجموعه‌های برچسب‌گذاری نشده چندگانه
۲۳۶.....	۵.۲.۲. اصول تئوری
۲۳۹.....	۵.۲.۳. ایجاد کلاس بندها: رهیافت دو-گانه
۲۴۸.....	۵.۲.۴. ایجاد کلاس بندها: رهیافت مستقیم
۲۵۱.....	۵.۲.۵. بحث و گفت و گو
۲۵۷.....	بخش دوم داده کاوی وب
۲۵۹.....	فصل ششم
۲۵۹.....	. بازیابی اطلاعات و جستجوی وب
۲۶۲.....	۶.۱. اصول اساسی در بازیابی اطلاعات
۲۶۶.....	۶.۲. مدل‌های بازیابی اطلاعات
۲۶۷.....	۶.۲.۱. مدل بولین
۲۶۸.....	۶.۲.۲. مدل فضای بردار
۲۷۲.....	۶.۲.۳. مدل زبان آماری
۲۷۴.....	۶.۳. بازخورد ارتباط

۶.۴. معیارهای ارزیابی	۲۷۸
۶.۵. پیش پردازش صفحه وب و متن	۲۸۳
۶.۵.۱. حذف stopword	۲۸۳
۶.۵.۲. ریشه یابی	۲۸۴
۶.۵.۳. وظایف دیگر مربوط به پیش پردازش برای متن	۲۸۵
۶.۵.۴. پیش پردازش صفحه وب	۲۸۶
۶.۵.۵. تشخیص (اسناد) دو گانه	۲۸۹
۶.۶. اندیس وارونه و فشرده سازی آن	۲۹۰
۶.۶.۱. اندیس وارونه	۲۹۱
۶.۶.۲. جستجو با استفاده از یک اندیس وارونه	۲۹۳
۶.۶.۳. ساختن اندیس	۲۹۴
۶.۶.۴. فشرده	۲۹۷
۶.۷. رابطه ی پنهان معنایی	۳۰۷
۶.۷.۱. تجزیه مقدار یکتایی	۳۰۸
۶.۷.۲. پرس وجو و بازیابی	۳۱۲
۶.۷.۳. یک مثال	۳۱۳
۶.۷.۴. بحث و گفتگو	۳۱۷
۶.۸. جستجوی وب	۳۱۸
۶.۹. ابر-جستجو و ترکیب چند رتبه بندی	۳۲۲

۳۲۴.....	۶.۹.۱. ترکیب با استفاده از امتیازات شباهت
۳۲۵.....	۶.۹.۲. ترکیب با استفاده از موقعیت رتبه‌ها
۳۲۹.....	۶.۱۰. اسپمینگ در وب
۳۳۰.....	۶.۱۰.۱. اسپمینگ محتوا
۳۳۲.....	۶.۱۰.۲. اسپمینگ لینک
۳۳۳.....	۶.۱۰.۳. تکنیک‌های مخفی‌سازی
۳۳۵.....	۶.۱۰.۴. مبارزه با اسپم
۳۳۹.....	فصل هفتم تحلیل لینک
۳۴۱.....	۷.۱. تحلیل شبکه اجتماعی
۳۴۲.....	۷.۱.۱. مرکزیت
۳۴۶.....	۷.۱.۲. اعتبار (نفوذ)
۳۴۹.....	۷.۲. co-citation (هم-استاد) و همبستگی فهرستی
۳۵۰.....	۷.۲.۱. co-citation
۳۵۱.....	۷.۲.۲. همبستگی فهرستی
۳۵۲.....	۷.۳. PageRank
۳۵۲.....	۷.۳.۱. الگوریتم PageRank
۳۶۳.....	۷.۳.۲. مزایا و معایب PageRank
۳۶۳.....	۷.۳.۳. PageRank زمانی
۳۶۵.....	۷.۴. HITS

۳۶۷.....	۷.۴.۱. الگوریتم HITS
۳۷۰.....	۷.۴.۲. پیدا کردن بردار مشخصه‌های دیگر
۳۷۱.....	۷.۴.۳. روابط با Co-Citation و زوج‌های کتاب‌شناختی
۳۷۲.....	۷.۴.۴. مزایا و معایب HITS
۳۷۴.....	۷.۵. تشخیص انجمن
۳۷۵.....	۷.۵.۱. تعریف مسئله
۳۷۷.....	۷.۵.۲. انجمن‌های دارای هسته‌ی دوبخشی
۳۸۱.....	۷.۵.۳. انجمن‌های با جریان ماکزیمم
۳۸۵.....	۷.۵.۴. انجمن‌های ایمیل مبتنی بر بینایی
۳۸۶.....	۷.۵.۵. انجمن‌های همپوشان از موجودیت‌های نام‌گذاری شده
۳۸۹.....	فصل هشتم خزش در وب
۳۹۱.....	۸.۱. یک الگوریتم خزنده پایه
۳۹۴.....	۸.۱.۱. خزنده‌های اول- سطح
۳۹۵.....	۸.۱.۲. خزنده‌های حرفه‌ای
۳۹۶.....	۸.۲. مسائل مربوط به پیاده‌سازی
۳۹۶.....	۸.۲.۱. واکنشی
۳۹۷.....	۸.۲.۲. تجزیه (تقسیم)
۳۹۹.....	۸.۲.۳. حذف کلمات اضافی و ریشه‌یابی
۳۹۹.....	۸.۲.۴. استخراج لینک و استانداردسازی (قانونی‌سازی)

۴۰۲.....	۸.۲.۵. تله‌های عتکبوت
۴۰۴.....	۸.۲.۶. مخزن صفحه
۴۰۵.....	۸.۲.۷. همزمانی
۴۰۷.....	۸.۳. خزنده‌های جهانی
۴۰۷.....	۸.۳.۱. مقیاس پذیری
۴۰۹.....	۸.۳.۲. پوشش؛ تازگی، و اهمیت
۴۱۱.....	۸.۴. خزنده‌های تمرکزی
۴۱۵.....	۸.۵. خزنده‌های موضوعی
۴۱۹.....	۸.۵.۱. محلیت موضوعی و سرنخ‌ها (نشانه‌ها)
۴۲۷.....	۸.۵.۲. نسخه‌های اول- بهترین
۴۳۱.....	۸.۵.۳. تطبیق
۴۴۰.....	۸.۶. ارزیابی
۴۴۶.....	۸.۷. کشمکش‌ها و مسائل اخلاقی در رابطه با خزنده‌ها
۴۵۱.....	۸.۸. بعضی از توسعه‌های جدید
۴۵۵.....	فصل نهم استخراج داده‌های ساختاریافته: تولید غلاف (بسته‌بند)
۴۵۶.....	۹. استخراج داده‌های ساختاریافته: تولید غلاف (بسته‌بند)
۴۵۷.....	۹.۱. مقدمات
۴۵۸.....	۹.۱.۱. دو نوع صفحه‌ای که سرشار (غنی) از داده هستند
۴۶۲.....	۹.۱.۲. مدل داده

۴۶۵.....	۹.۱.۳. کد گذاری HTML نمونه‌های داده
۴۶۶	۹.۲. القا بسته‌بند (غلاف)
۴۶۸.....	۹.۲.۱. استخراج از یک صفحه
۴۷۲.....	۹.۲.۲. یادگیری قانون استخراج
۴۷۸.....	۹.۲.۳. شناسایی نمونه‌های اطلاعاتی
۴۷۹.....	۹.۲.۴. نگهداری بسته‌بند
۴۸۰.....	۹.۳. یادگیری بسته‌بند مبتنی بر نمونه
۴۸۵.....	۹.۴. تولید خودکار بسته‌بند: مسائل و مشکلات
۴۸۶.....	۹.۴.۱. دو مشکل استخراج
۴۸۷.....	۹.۴.۲. الگوها به عنوان عبارات منظم
۴۸۸.....	۹.۵. مطابقت رشته و مطابقت درخت
۴۸۹.....	۹.۵.۱. فاصله ویرایش رشته
۴۹۲.....	۹.۵.۲. مطابقت درخت
۴۹۶	۹.۶. تنظیم چندگانه
۴۹۶.....	۹.۶.۱. روش ستاره مرکزی
۴۹۹.....	۹.۶.۲. تنظیم درخت جزئی
۵۰۵.....	۹.۷. ساختن درخت‌های DOM
۵۰۷.....	۹.۸. استخراج بر اساس یک صفحه لیست واحد: رکوردهای داده مسطح
۵۰۸.....	۹.۸.۱. دو ملاحظه درباره رکوردهای داده

۵۱۱.....	۹.۸.۲. ماینینگ نواحی داده.....
۵۱۸.....	۹.۸.۳. شناسایی رکوردهای داده در یک ناحیه داده.....
۵۱۹.....	۹.۸.۴. تنظیم و استخراج آیتم داده.....
۵۲۰.....	۹.۸.۵. استفاده از اطلاعات بصری.....
۵۲۱.....	۹.۸.۶. برخی تکنیک‌های دیگر.....
۵۲۲.....	۹.۹. استخراج بر اساس یک صفحه لیست واحد: رکوردهای داده تودرتو.....
۵۳۱.....	۹.۱۰. استخراج بر اساس چندین صفحه.....
۵۳۱.....	۹.۱۰.۱. استفاده از تکنیک‌های بخش‌های قبلی.....
۵۳۲.....	۹.۱۰.۲. الگوریتم RoadRunner (کوکوسان).....
۵۳۴.....	۹.۱۱. برخی مسائل دیگر.....
۵۳۴.....	۹.۱۱.۱. استخراج از صفحات دیگر.....
۵۳۵.....	۹.۱۱.۲. ترکیب فصلی یا اختیاری.....
۵۳۷.....	۹.۱۱.۳. یک نوع مجموعه یا یک نوع چندتایی.....
۵۳۷.....	۹.۱۱.۴. برچسب‌گذاری و یکپارچه‌سازی.....
۵۳۸.....	۹.۱۱.۵. استخراج خاص دامنه.....
۵۳۹.....	۹.۱۲. بحث و گفتگو.....
۵۴۱.....	فصل دهم یکپارچه‌سازی اطلاعات.....
۵۴۳.....	۱۰.۱. مقدمه ای بر مطابقت شما.....
۵۴۶.....	۱۰.۲. پیش‌پردازش برای تطبیق شما.....

۵۴۷.....	۱۰.۳. مطابقت سطح شما
۵۴۸.....	۱۰.۳.۱. رهیافت‌های زبانی
۵۵۰.....	۱۰.۳.۲. رهیافت‌های مبتنی بر شروط (قیود)
۵۵۱.....	۱۰.۴. مطابقت سطح نمونه و دامنه
۵۵۴.....	۱۰.۵. ترکیب شباهت‌ها
۵۵۶.....	۱۰.۶. تطبیق m_1 :
۵۵۸.....	۱۰.۷. برخی مسائل دیگر
۵۵۸.....	۱۰.۷.۱. استفاده مجدد از نتایج تطبیق قبلی
۵۵۹.....	۱۰.۷.۲. تطبیق تعداد زیادی از شماها
۵۶۰.....	۱۰.۷.۳. نتایج تطبیق شما
۵۶۰.....	۱۰.۷.۴. تعاملات کاربر
۵۶۱.....	۱۰.۸. یکپارچه‌سازی تعاملات پرس‌وجوی وب
۵۶۵.....	۱۰.۸.۱. یک رهیافت مبتنی بر خوشه‌بندی
۵۶۹.....	۱۰.۸.۲. یک رهیافت مبتنی بر همبستگی
۵۷۸.....	۱۰.۹. ساختن یک رابط پرس‌وجوی عمومی یکپارچه
۵۷۹.....	۱۰.۹.۱. تناسب ساختاری و الگوریتم ادغام
۵۸۱.....	۱۰.۹.۲. تناسب واژه ای
۵۸۲.....	۱۰.۹.۳. تناسب نمونه
۵۸۴.....	منابع

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 به اختصار وب) تقریباً بر هر جنبه‌ای از زندگی روزمره ما تأثیر عمیقی داشته است. این شبکه بزرگ‌ترین و شناخته‌شده‌ترین منبع اطلاعاتی است که به راحتی قابل دسترسی و جستجو می‌باشد. شبکه وب از اسناد به هم پیوسته 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 به راحتی از خانه یا دفتر کار خود اطلاعات مورد نیاز را پیدا کرده و دانش خود را با دیگران به اشتراک بگذاریم.

وب همچنین به یک کانال مهم برای انجام کسب و کار 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 به عنوان "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 ذخیره شده 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 اسکپ، 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^۱ 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، برنرزی دوباره پیشنهاد خود را مطرح کرد و این بار حمایت لازم برای شروع کار را به دست آورد. با این پروژه، برنرزی و تیم او در CERN پایه‌های توسعه آینده وب به‌عنوان یک سیستم ابرمتن توزیع‌شده را بنا نهادند. آنها سرور و مرورگر خود، پروتکل ارتباطی بین کلاینت و سرور یعنی پروتکل انتقال ابرمتن (HTTP)، زبان نشانه‌گذاری ابرمتن (HTML) برای نگارش اسناد وب، و آدرس‌های یکتای منبع (URL) را معرفی کردند. و این‌گونه بود که وب آغاز شد.

ظهور مرورگرهای موزاییک^۱ و نت اسکپ: 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 به سرعت 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 اسکپ کامیونیکیشنز تغییر نام داد) را تأسیس کردند. در عرض چند ماه، مرورگر نت اسکپ 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 اسکپ پرداخت. ایجاد شبکه جهانی وب توسط تیم برنرزی 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 واشنگتن دی‌سی به نمایش گذاشته شد. در این کنفرانس، دانشمندان آرپا کامپیوترهایی از ۴۰ نقطه مختلف را به هم متصل کردند. در سال ۱۹۷۳، ویتون 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. باین‌حال، آرپانت همچنان به عنوان 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 به عنوان 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 به عنوان 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 به سرعت رشد کرد و سرمایه‌گذاران با دیدن فرصت‌های تجاری، به آن علاقه‌مند شدند. کسب و کارهای 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 قابل توجهی را تجربه کرده است و خود را به عنوان 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 با ارزش 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؛ به طوری که 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 قابل توجهی 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 به عنوان 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 به عنوان صفحات با کیفیت 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 حق تألیف و سیاست‌های حفظ حریم خصوصی. برای یک کاربرد خاص، تنها بخشی از این اطلاعات مفید است و بقیه به عنوان 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 کسب و کار 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 به راحتی 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 داده کاوی 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 [۲۶، ۱۵] مراجعه کنید. برای بررسی استخراج اطلاعات از وب، لازم است با داده کاوی آشنا شویم، چرا که این تکنیک در بسیاری از وظایف مرتبط با استخراج از وب مورد استفاده قرار گرفته است. با این حال، استخراج اطلاعات از وب صرفاً به کارگیری داده کاوی 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 داده کاوی چیست؟

داده کاوی یا به عبارتی استخراج داده‌ها که به آن کشف دانش در پایگاه‌های داده (KDD) نیز گفته می‌شود، به‌عنوان فرایند کشف الگوها یا دانش مفید از منابع مختلف داده؛ مانند پایگاه‌های داده، متون، تصاویر و وب شناخته می‌شود. الگوهایی که استخراج می‌شوند باید معتبر، بالقوه مفید و قابل درک 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 یا دانش 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 به صورت تکراری انجام می‌شود و ممکن است چندین بار تکرار شود تا به نتیجه نهایی و رضایت بخش دست یابد که سپس در وظایف عملیاتی دنیای واقعی مورد استفاده 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 (وب کاوی) یا استخراج داده از وب به منظور کشف اطلاعات مفید یادانش از ساختار لینک‌ها، محتوای صفحات وب و داده‌های مربوط به استفاده کاربران انجام می‌شود. اگرچه وب ماینینگ از تکنیک‌های مختلف داده کاوی بهره می‌برد، اما به دلیل ناهمگونی و طبیعت نیمه ساختاریافته یا غیرساختاریافته داده‌های وب، نمی‌توان آن را به طور کامل به عنوان یک کاربرد ساده از تکنیک‌های سنتی داده کاوی 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 مورد استفاده 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. به عنوان مثال، از طریق لینک‌ها می‌توان صفحات وب مهم را شناسایی کرد، که این فناوری در موتورهای جستجو مورد استفاده قرار می‌گیرد. همچنین می‌توان جوامع کاربری با علاقه‌مندی‌های مشترک را کشف کرد. این وظایف در داده کاوی 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. به عنوان مثال، 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. با این حال، 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 سه مرحله‌ای مشابهی را دنبال می‌کنیم: پیش‌پردازش داده‌ها، داده‌کاوی وب و پس‌پردازش. با این حال، روش‌هایی که در هر یک از این مراحل به کار گرفته می‌شوند، ممکن است تفاوت قابل توجهی با تکنیک‌های مورداستفاده در داده‌کاوی سنتی داشته باشند.

۱.۴ خلاصه فصل‌ها

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. از آنجایی که مرز دقیقی بین جستجوی وب و وب کاوی 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 داده کاوی می‌پردازد که در بسیاری از وظایف وب کاوی، به ویژه 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 داده کاوی و وب کاوی است. در این روش، از داده‌هایی با کلاس‌های از پیش تعریف شده برای یادگیری یک تابع طبقه‌بندی استفاده می‌شود که می‌تواند برای طبقه‌بندی داده‌های جدید مورد استفاده قرار گیرد.

فصل ۴ به یادگیری بدون نظارت می‌پردازد، که در آن داده‌ها بدون برچسب‌های از پیش تعریف شده 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 به منظور 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 مورد نیاز 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، به عنوان یک ویژگی خاص در تحلیل شبکه‌های اجتماعی مورد استفاده 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 به صورت خودکار ساختار هایپرلینک‌های وب را طی کرده و صفحات وب را به صورت 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 ارزش افزوده مانند خرید مقایسه‌ای و جستجوی فراگیر را فراهم می‌کند.

فصل ۱۰ به یکپارچه‌سازی اطلاعات می‌پردازد. برای استفاده از داده‌ها یا اطلاعات استخراج شده از چندین سایت به منظور ارائه خدمات ارزش افزوده، باید این داده‌ها/اطلاعات را به صورت معنایی یکپارچه کنیم تا یک پایگاه داده 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 شخصی‌سازی شده برای کاربران ایجاد کرد.

۱.۵ نحوه مطالعه کتاب