شالودهٔ داده‌کاوی و کشف دانش

● **فریبرز درودی**
دانشجوی دکتری کتابداری و اطلاع‌رسانی دانشگاه آزاد اسلامی، واحد علوم و تحقیقات تهران

■ **Foundations of Data Mining and Knowledge Discovery.** Tsau Young Lin ...[et al.]. Berlin; NewYork : Springer-Verlag.
ISBN: 978-3-540-26257-2

تسو یانگ لین[۱] سرویراستار اثر حاضر، استاد علوم رایانه در دانشگاه ایالتی سن ژوزه[۲] است. وی متخصص علوم رایانه و در حوزه‌های مختلف صاحب پیشینهٔ آموزشی و پژوهشی است. زمینه‌های حرفه‌ای که لین بدانها پرداخته، عبارت‌اند از: پایگاه‌های داده، داده‌کاوی[۳]، امنیت داده[۴]، نظام‌های داده و دانش‌مدار[۵]، محاسبات فازی[۵]، کنترل هوشمند[۶] و مهندسی الکترونیک[۷]. وی با ارائهٔ ۲۱۵ مقالهٔ علمی در مجلات تخصصی و همایش‌ها، و نیز تدوین ۱۱ کتاب، کارنامه‌ای پربار از خود برجای نهاده است (صفحهٔ خانگی تسو یانگ لین[۸]). چهار ویراستار دیگر کتاب نیز از متخصصان علوم رایانه‌اند که در زمینهٔ فعالیت حرفه‌ای داده‌کاوی و پایگاه‌های داده، و نیز نظام‌های اکتشافی دانش صاحب آثار مهم و قابل توجهی‌اند.

داده‌کاوی یکی از زمینه‌های مهم دربارهٔ موضوع پایگاه داده است. با افزایش روزافزون منابع اطلاعاتی و داده‌های علمی و ورود آنها به پایگاه‌های اطلاعاتی، حجم داده‌ها افزایش چشمگیری یافته، به‌گونه‌ای که بازیابی اطلاعات مرتبط در درون این پایگاه‌ها با دشواری‌های خاص خود روبه‌رو شده و دستیابی به اطلاعات ظرافت بالایی می‌طلبد. نظریات داده‌کاوی و بازیابی داده‌های مرتبط، در کنار فنون و الگوهای انجام این فرآیند، حوزه‌ای از مباحث نظری و کاربردی را مطرح ساخته که امروزه در طراحی، تدوین و مدیریت پایگاه داده بسیار مؤثر است. همچنین با توسعهٔ پایگاه‌های دانش[۹] و بهره‌گیری از آن در سازمان‌ها و مؤسسات مختلف، بحث اکتشاف دانش و آشنایی با ساختارهای اثربخش آن، اهمیت بسیاری یافته است. بر این اساس امروزه موضوع داده‌کاوی در کنار اکتشاف دانش، از جمله مباحثی است که در میان متخصصان علوم رایانه با استقبال کم‌نظیری روبه‌رو شده است و کتاب حاضر که حاصل مقالات مهمی در این حوزهٔ مطالعاتی است، اخیراً به چاپ رسیده است.

این اثر در حوزهٔ مطالعاتی هوش محاسباتی و ششمین مجلد از این فروست است که یافته‌های نوین و حوزه‌های تخصصی گوناگونی را در این عرصه پوشش می‌دهد. کتاب حاوی مقالات ارزشمندی است و تلاش نویسندگان در معرفی و بررسی آخرین یافته‌های پژوهشی در عرصهٔ کلی هوش محاسباتی، که به حوزهٔ تخصصی داده‌کاوی و اکتشاف دانش ختم شده، درخور ستایش است. از نکات برجستهٔ دیگر باید به ترکیب مناسب میان نظریات، کاربردها و طراحی روش‌های علمی اشاره کرد که جملگی در این عرصه به‌شکل مطلوبی بیان شده است.

برخی از مباحث کتاب در زمینهٔ دانش رایانه است که با مباحث مهندسی و مکانیک نیز ارتباط دارد. استفاده از علوم مختلف در ایجاد فضایی مناسب برای ارائهٔ یافته‌های نوین پژوهشی نشان‌دهندهٔ توانمندی‌های بالای نویسندگان است. مطالب این اثر در واقع مقالات ارائه‌شده در یک کارگاه آموزشی است که ششم می ۲۰۰۲ در تایپه[۱۰] مرکز تایوان[۱۱] برگزار شد و بخشی از ششمین همایش آسیا- اقیانوسیه در موضوع اکتشاف دانش و داده‌کاوی[۱۲] بود. کتاب در سه بخش تدوین شده است. بخش اول، «مبانی داده‌کاوی»، هشت مقاله را دربرمی‌گیرد. بخش دوم، «روش‌های داده‌کاوی» شامل نُه مقاله است و بخش سوم «کشف دانش عمومی» از چهار مقاله تشکیل شده است.

مقالهٔ اول دربارهٔ موضوع اکتشاف دانش به‌مثابهٔ ترجمه است. در این بخش، مروری بر تسخیر اکتشافی دانش به‌عنوان ترجمه از ارائهٔ غیرنمادین[۱۳] به نمایش نمادین[۱۴] شده است. هر دو مفهوم در چارچوب مفاهیم اکتشافی دانش به‌روشنی مورد توجه قرار گرفته و ویژگی‌های نمایش نمادین در قالب سنجش‌های کمی مطرح شده است. مراحل پردازش ارائهٔ غیرنمادین ویژگی‌های مغایر با نوع نمادین دارد؛ بنابراین شکاف و فاصلهٔ عمیقی میان آنها وجود داشته که در این مقاله تلاش

با توسعهٔ پایگاه‌های دانش و بهره‌گیری از آن در سازمان‌ها و مؤسسات مختلف، بحث اکتشاف دانش و آشنایی با ساختارهای اثربخش آن، اهمیت بسیاری یافته است

شده اســت تا با ارائهٔ سنجشــی کمی و روش‌مند، ابعاد و خصوصیات آن معرفی شود. مقالهٔ دوم به مبانی ریاضی قوانین پیوند و پیوندکاوی با شــیوهٔ حل نامعادله‌های خطی انتگرال اختصاص دارد. نویســنده در ایــن مقاله با بیان الگوهای ریاضی قوانین پیوند می‌کوشــد تا رویکرد نوینی از پیوندکاوی ارائه دهد. به‌طور غیررسمی، داده‌کاوی از الگوهای برآمده از داده مشــتق می‌شــود. مکانیک ریاضی کاوش پیوند از این دیدگاه، به‌دقت مورد بررســی و آزمایش قرار گرفته است. دادهٔ جدولی از نمادهاست و الگوی هر تجربهٔ منطقی- جبری بوده که از این جدول با پشــتیبانی زیاد منشــأ گرفته است. علاوه بر آن مدلی از داده در این مقاله ارائه شده که به این فرآیند یاری می‌رساند.

در مقالهٔ ســوم، «ارائهٔ مطالعه‌ای تطبیقی دربارهٔ مدل‌های کاوش الگوی متوالی»، مشــکلات مربوط به الگوهای متوالی کاوش بررسی شده و با بهره‌گیری از یک روش ارزشیابی عمومی، ارزش کیفی تعیین شده است. در این پژوهش چهار مؤلفهٔ اصلی ارزشیابی مطرح و تحلیل دقیق شده است. این مطالعه همچنین با سنجش وضع پایگاه‌های دادهٔ ترکیبی و بهره‌گیری از روش پیمایشــی به شــیوهٔ تصادفی وضع آنان را بررســی و ســطوح اختلال ایجاد شــده را معرفی می‌کند. همچنین راه‌حل‌هــای احتمالــی و قریب به واقع معرفی و مــدل مورد نظر بیان می‌شــود. نویســندگان در مقالهٔ چهارم با معرفی مدل‌های همبستگی توانمند به توضیح بیشتر دربارهٔ کاربرد آنها می‌پردازند. کانون توجه این مطالعه بر ترجیح‌هایی است که نسبت به مدل‌های رقابتی وجود دارد. آمار کلاسیک، بخشــی از فنون شناخته شده برای انتخاب مدل‌هایی در رگرسیون‌های چندجمله‌ای است؛ که در این مقاله مورد مداقه قرار گرفته است. همچنین اصول انتخاب مدل با ویژگی‌های خاص بررسی و نتایج مهمی از آن به‌دست آمده است.

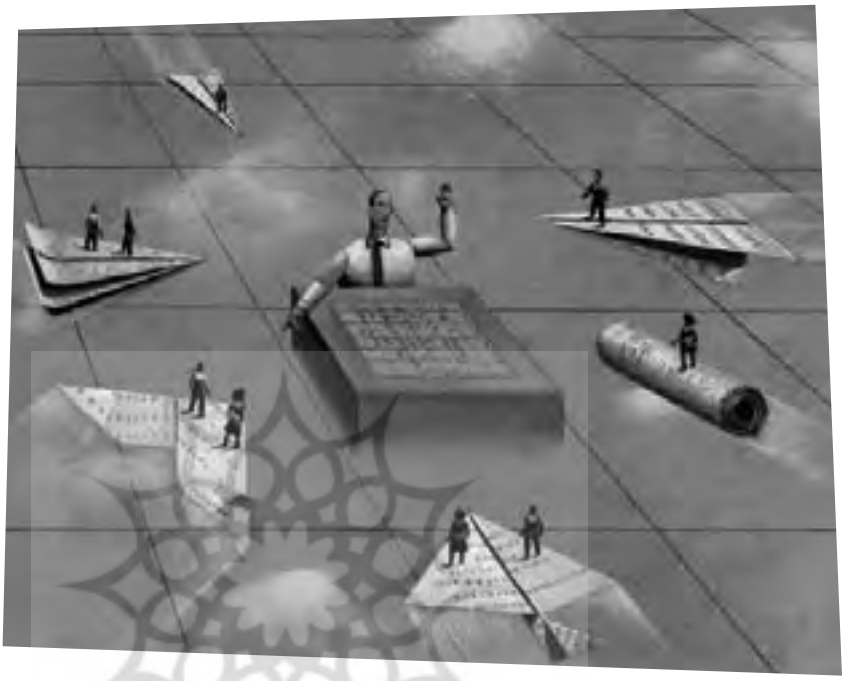
مقالهٔ پنجــم نیز به چارچوب‌هــای احتمالی منطقی‌مــدار دربارهٔ ویژگی‌های اکتشــاف دانش در پایگاه‌های داده[15] می‌پردازد. به‌منظور ارتقاء و توســعهٔ فرآیند پردازشی اکتشــاف دانش در پایگاه‌های داده، انواع شیوه‌های اکتشاف دانش بررسی می‌شود. در این مقاله بنیان‌های منطقی در فرآیند اکتشــاف دانش تحلیل شده و برخی از ویژگی‌های مهم آن مطرح می‌شــود. نویسندگان پس از بررسی‌ها، منطق احتمالی باچیوس[16] را گزینهٔ مطلوب و مناسبی در فعالیت اکتشاف دانش معرفی می‌کنند. در ادامه بر اســاس زبان خــاص این منطق احتمالی، الگویی ارائه می‌شود که پیش از آن مطرح نشده، و در اکتشاف دانش سودمند لســت. مقالهٔ ششــم دربارهٔ نگرش دقیق در اســتفاده از روش‌شناسی آماری در داده‌کاوی است. اکتشاف دانش در پایگاه‌های داده به‌صورت ذاتی یک فعالیت آماری است. بسیاری از روش‌شناسی‌های کلاسیک آماری برای اهدافی که هریک می‌توانند بســیار متفاوت از روش‌های اکتشاف دانش در پایگاه‌های داده باشند، طراحی شده‌اند. در این مقاله بیان شــده که حوزهٔ اکتشاف دانش در پایگاه‌های داده از روش‌شناسی آمار استفادهٔ بسیار مفیدی می‌کند.

مقالــهٔ هفتم به موضــوع انتخاب فرضیه و تأیید درســتی آن در داده‌کاوی اختصاص دارد. نویســندگان در این مقاله با بیان این مطلب که داده‌کاوی نمونه‌ای از یک روش‌شناسی استقرائی[17] است، به تبیین ملاحظــات فلســفی و معرفتــی آن می‌پردازند. به‌طور خــاص، اثبات صحت و درســتی روش اســتقرائی یک مشــکل تاریخــی در عرصهٔ مباحث معرفت‌شناســی اســت که تلاش برای فرمول‌بنــدی و تبیین دقیق ویژگی‌های آن فعالیت مهمی محســوب می‌شــود. مقالهٔ هشتم به بررســی اســتقلال آماری دربارهٔ جدول احتمالی می‌پردازد. در این مقاله تحلیل مناسبی از وضع استقلال آماری با توجه به منطق ریاضی ارائه شــده که در قالب جدول‌های احتمالی و تأثیر بر کاربرد آن مورد بررسی قرار گرفته است.

در مقالهٔ نهم پژوهشــی تطبیقی درخصوص انتخاب مدل به‌شیوهٔ تحلیل عامل صفر و یک انجام شده که در مباحث تحلیل داده کاربرد وسیع و متعددی دارد. این بررسی‌ها بیشتر دربارهٔ تبیین مدلی است که بتواند در فرآیندهای تحلیل داده و بررســی‌های مربوط به آن به یاری متخصصان آید. علاوه بر آن تلاشی درخور انجام پذیرفته تا معیارهای اصلی در انتخاب مدل معرفی شــده و بر مبنای محاسبات آماری ارتقا داده شود. در مقالهٔ دهم دربارهٔ مسئلهٔ استخراج قوانین عمومی مربوط به ویژگی‌های انتزاع خودکارسازی‌شده و استفاده از روش ناول[18] برای قوانین عمومی‌شــدهٔ کاوش به‌همراه پشتیبانی قوی از ساختارهای آن بحث می‌شود.

مقالهٔ یازدهم به بررســی فرآینــد تصمیم‌گیری بــر مبنای نظام دوگانه[19] در دانش‌های چندگانه اختصاص دارد. تصمیم‌گیری مبتنی بر منطق ریاضی و با بهره‌گیری از روش‌های آماری یکی از فعالیت‌های خاص داده‌کاوی اســت که در این مقاله با معرفی مدلی ویژه ابعاد آن بیشتر تحلیل شده است. در مقالهٔ دوازدهم مباحثی دربارهٔ ویژگی‌های صورت‌گرایی منطق‌مدار برای داده‌کاوی موقتی مطرح و روش‌شناسی خاص آن برای اکتشــاف دانش بیان شده است. از جمله مسائل مورد توجه در این مقاله، مفهوم ســاختار زمانی خطی اســت که در قوانین داده‌کاوی کاربرد دارد.

مقالهٔ ســیزدهم به موضوع رویکرد جانشین در قوانین پیوندکاوی اختصــاص دارد که پایه و مبنای آن در ارائهٔ داده‌های تحلیلی بوده که از طریق رشــته‌های مناســبی از واحد بیت[20] تبیین شــده است. برای



انتخاب عنوان‌های هر مقاله و نیز عنوان‌های فرعی در ساختار مقاله بر مبنای محتوای اثر صورت پذیرفته و ارتباط عمیقی میان آنها وجود دارد

انواع متعددی از قوانین پیوند شـامل قوانیـن پیوند موقعیتی می‌توان از آن بهـره گرفت. مقالهٔ چهاردهم به کاوش مسـتقیم قوانین از داده می‌پـردازد. چگونگی کسـب قوانین مهم این حـوزه از طریق تجزیه منطقـی داده به فرآیند مطالعاتی خاصی نیاز دارد که نویسـندگان این مقاله به تبیین آن پرداخته‌اند. همچنین موضوع از دسـت دادن مقادیر از منظر ریاضی مطرح و بررسی می‌شود.

مقالهٔ پانزدهم دربارهٔ مبحث تعیین شناسـایی خوشـه‌ها[۲۱] است. پدیدهٔ خوشـه یکی از مباحث عمده‌ای است که در عرصهٔ داده‌کاوی و اکتشـاف دانش مطرح می‌شود و کاربرد وسـیعی در نظام‌های بازیابی اطلاعات دارد. در این مقاله همچنین بهره‌گیری حداکثری از پیکربندی آنتروپی[۲۲] مطرح شـده و دربارهٔ ویژگی‌های آن مطالب سودمندی ارائه شده است. در مقالهٔ شانزدهم الگوهای کاوش در شبکه‌های عصبی با سـاختار خاص آن و عناصر تأثیرگذار در فرآیند داده‌کاوی مورد بحث و بررسی قرار می‌گیرد.

مقالهٔ هفدهم شامل مبحث دانش‌کاوی توسعه‌یافته است. در این مقاله با ارائهٔ رویکردی چندروشـی، فرآیند کاری دانش‌کاوی و عوامل مهم دخیل در این شـیوه بررسی شده است. مقالهٔ هجدهم به موضوع آموزش بر اسـاس شیوهٔ ارسـال و بهره‌گیری از روش برچسب‌زنی از مجاری پیوسته اختصاص دارد. در این مقاله سیاهه‌ای از برچسب‌های قابل ارسال در فرآیند تعاملی معرفی و تحلیل شده است.

مقالـهٔ نوزدهم نیز دربارهٔ شناسـایی مقادیر تحلیلی اسـت که در ارتباط با نمایه‌سـازی معنایی راکد[۲۳] بیان می‌شـود. روش نمایه‌سازی معنایـی یکی از شـیوه‌های مفیدی اسـت که در فعالیـت داده‌کاوی اسـتفاده شده و مبتنی بر تحلیل آماری است. نویسندگان در این مقاله با بهره‌گیری از روش‌هـای ریاضی به تبیین این مقوله می‌پردازند. در مقالهٔ بیستم موضوع گزارش‌های نتایج داده‌کاوی در قالب زبان طبیعی مطرح و ابعاد خاص آن بررسـی می‌شـود. همچنین شـیوه‌های نوین سنجش گزارش‌ها و ارائهٔ نتایج در محیط‌های دیجیتالی با بهره‌گیری از زبان طبیعی به‌تفصیل بیان شده است. مقالهٔ بیست‌ویکم به الگوریتم محاسـباتی مقدار محتمل دربارهٔ افزایـش کنترل کار کاربر اختصاص دارد. در ایـن مقاله شـیوه‌های نظـارت بر فعالیـت کاری کاربران در فعالیـت داده‌کاوی در پایگاه‌های اطلاعاتی با تکیه بر الگوریتم خاص محاسباتی مقدار محتمل بررسی و نتایج آن ارائه شده است.

کتـاب مبانی داده‌کاوی و کشـف دانش شـامل آخریـن نتایج و راهنمایی‌های نوین در زمینهٔ پژوهش‌های داده‌کاوی است. داده‌کاوی که بـا فناوری‌های متعددی مرتبط بوده، شـامل مباحثی چون هوش رایانه‌ای،[۲۴] پایگاه داده و مدیریت دانش،[۲۵] آموزش ماشینی،[۲۶] محاسبهٔ تقریبی،[۲۷] و آمار اسـت که از جمله حوزه‌هایی‌اند که سرعت بالایی در رشـد و توسعهٔ علوم رایانه دارند. این کتاب به ارائهٔ نتایج پژوهش‌های انجام شده، دربارهٔ بنیادهای رشتهٔ علمی و ارائهٔ موقعیت پیشرفته‌ترین فناوری برای بسـیاری از پژوهش‌های جاری می‌پردازد. علاوه بر آن اثر حاضر به اثبات ارزش و سودمندی داده‌کاوی برای پژوهشگران نیز می‌پردازد. به‌گونه‌ای که اگر آنان مایل نباشند اصول بنیادی داده‌کاوی را در پژوهش‌های خود پوشـش دهند یا به نظریات کاربردهای عملی توجه داشته باشـند، در هر دو صورت می‌توانند از آن استفادهٔ مناسبی داشته باشند (سایت آمازون).

مطالب کتاب، از مباحث نوین حوزهٔ مطالعاتی داده‌کاوی و اکتشاف دانش است. این مباحث در قالب مقالات مستقل در مجلات علمی به

چاپ رسیده یا در همایش ارائه شده‌اند. علاوه بر آن در برخی از متون درسی از آنها استفاده شده یا مقالات در قالب درس‌های تخصصی نیز ارائه شده‌اند. اهمیت موضوع در حد بالایی است و نویسندگان نیز با پژوهش‌های کاربردی و پیمایشی به غنای اثر افزوده‌اند. یکی از ویژگی‌های مهم کتاب برقراری ارتباط منطقی میان روش‌های داده‌کاوی و اکتشاف دانش است. الگوهای برجسته‌ای که توانمندی بازیابی اطلاعات را افزایش می‌دهد، حاصل تجربه‌های سودمندی است که در قالب نگارش‌های پژوهشی به‌دست آمده است. از همین‌رو برخی از الگوریتم‌های مناسب برای طراحی الگوهای مطلوب معرفی و تحلیل شده است.

معرفی فناوری‌های نوین حوزهٔ داده‌کاوی و اکتشاف دانش نیز یکی دیگر از نقاط مثبت اثر است. برخی از این فناوری‌ها در مراحل نوین فعالیت خود هستند و می‌توانند مورد بررسی و تحلیل بیشتر قرار گیرند. از سوی دیگر توضیحات ارائه شده در کتاب کاملاً فنی و تخصصی است و به‌زبان ریاضی بیان شده است. در واقع رویکرد اصلی در بیان مطالب، قوانین و الگوهای ریاضی بوده که تا حد زیادی برای خوانندگان آشنا با دانش ریاضیات و رایانه سودمند است. الگوهایی که در اثر بررسی شده، دربارهٔ ساختارهای کلی قوانین ریاضی شرح داده شده و مدل‌هایی که مورد تحلیل قرار گرفته است، از همین شیوه تبعیت می‌کند. از مزایای اثر به ارائهٔ نتایج نوینی که از بررسی‌ها و مطالعات پیمایشی به‌دست آمده، می‌توان اشاره کرد که برخی از آنها کاملاً جدید بوده و در پژوهش مورد نظر به‌دست آمده است. الگوهای ارائه شده بیشتر با روش‌شناسی منطق ریاضی بیان شده و این یکی از نقاط مثبت در ساختار اثر محسوب می‌شود.

ارائهٔ چکیدهٔ مباحث در ابتدای هر مقاله سبب شده تا خواننده بتواند در آغاز مطالعهٔ خود از محتوای مقاله به‌خوبی آگاهی یافته و با کلیات بحث آشنا شود. در برخی از مقالات شیوه‌های برجسته‌سازی و نمایش اطلاعات و نتایج مهم به‌خوبی رعایت و اهمیت بحث مطرح شده است. انتخاب عنوان‌های هر مقاله و نیز عنوان‌های فرعی در ساختار مقاله بر مبنای محتوای اثر صورت پذیرفته و ارتباط عمیقی میان آنها وجود دارد. منابع کتاب‌شناسی هر مقاله در پایان آن آورده شده است. این اثر را می‌توان یکی از منابع مهم در زمینهٔ داده‌کاوی و اکتشاف دانش به‌شمار آورد.

پی‌نوشت‌ها

1. Tsau Young Lin
2. San Jose State University
3. Data Mining
4. Data Security
5. Data and Knowledge Based System
6. Fuzzy Computing
7. Intelligent Control
8. E-Engineering
9. Tsau Young Lin» home page
10. Knowledge base
11. Taipei
12. Taiwan
13. 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining
14. Non-symbolic representation
15. symbolic representation
16. Knowledge Discovery in Databases(KDD)
17. Bacchus» probability logic
18. Inductive methodology
19. Novel method
20. Hybrid
21. Bit
22. Clusters
23. Entropy
24. Latent semantic indexing
25. Computational intelligence
26. Database and knowledge management
27. Machine learning
28. Soft computing

منابع و مآخذ

1. Amazon [web site], **Foundations of Data Mining and Knowledge Discovery.** Tsau Young Lin ... [et al.]. Editorial Reviews by Amazon.[on-line] Available:
http://www.amazon.com/Foundations-Knowledge-Discovery-Computational-Intelligence/dp/3540262571/ref=sr_1_7274500-9532555?ie=UTF8&s= books&qid=1191833211&sr=1-8[accessed8, , Oct., 2007].
2.Tsau Young Lin' home pag. [on-line]. Available: http://www.cs.sjsu.edu/~tylin/index.php?goto=index.php [accessed 9 , Oct., 2007]