کشف دانش در کتابخانه های رقومی *

نوشته: ت. ویشواناتان، زرین تاج برنایی، ر.ج. گوپتا
ترجمه: محسن عزیزی

چکیده

کتابخانه های سراسر دنیا دستخوش یک فرآیند دگرگونی هستند. کتابخانه های سنتی که عمدتا" دارای مواد چاپی هستند سریعا" در حال حرکت به سوی خودکارکردن فعالیت های خود از قبیل امانت و فهرست ماشین خوان می باشند. کتابخانه های خود کار نیز به سوی کتابخانه های الکترونیکی دارای منابع دیسک فشرده، و دسترسی به اینترنت، گام برمی دارند. کتابخانه های الکترونیکی که مدتی است از فعالیت آنها می گذرد در حال بررسی تغییر جهت به سمت کتابخانه های رقومی می گردد و از طریق شبکه در دسترس قرار می گیرد. کتابخانه های رقومی دارای مجموعه داده های بسیار گسترده ای به شکل رقومی هستند و برای ذخیرهٔسازمان یافتهٔ داده هایشان باید از مفاهیم انبارش داده ها استفاده کنند. کتابخانه های رقومی اساسا" تالارهای بزرگ دانش الکترونیکی هستند و بنابراین، کشف دانش در کتابخانه های رقومی (KDDL) به موضوع مهم بدل می شود. استخراج داده ها شاخه ای از دانش است که به کشف دانش پنهان، الگوهای نامنتظره، و قواعد جدید از داده پایگاه های الکترونیکی بزرگ می پردازد. رویکرد طبیعی به کشف دانش در کتابخانه های رقومی، بهره گیری از فنون موجود استخراج داده ها است. اما تلاش های کنونی در امر استخراج داده ها عمدتا" رو به کاربردهایی دارد که با داده های مالی، پایگاه های بزرگ خریداران، رکوردهای سرویس راهنمای کمکی، و . . . . مرتبط اند. ویژگی این کاربردها، داده های مرتبط و سازمان یافته، در مقیاس وسیع است. برعکس، تالارهای دانش کتابخانه های رقومی دارای مقادیر بسیاری از داده های سازمان نیافته در حوزه های موضوعی مختلف هستند. در نتیجه،فنون موجود در استخراج داده ها برای کشف دانش در کتابخانه های رقومی چندان مطلوب و مناسب نیست. این مقاله پس از بحث دربارهٔ فرآیند دگرگونی کتابخانه ها و نامناسب بودن فنون موجود در استخراج داده ها برای کشف دانش در کتابخانه های رقومی، برای این کار فنون جدیدی را بر مبنای مفاهیم سازماندهی دانش، که در کتابخانه ها استفادهٔ گسترده ای از آنها می شود، پیشنهاد می کند.

بسیاری از متفکران بزرگ، دنیا را در زمان کنونی در حال گذار می دانند. اشخاص مشهور در حوزه های مربوط به خود نظام های اقتصادی، سیاسی،و اجتماعی جدیدی را پیش بینی می کنند. پیش بینی می شود که در آیندهٔ نزدیک، نوعی نظم نوین جهانی برقرار گردد که برای یکی دو هزاره ماندگار باشد. صنایع متعارف راه را برای صنایع جدیدی که نقش مهمی در

× × - Vishwanathan, T; Bornaee, Zarrintaj & Gupta, R.G.(1998); "Knowkedge discovery In digital librabies (KDDL) "; in 49 thFID conference and congress, Jaipur & New Delhi, 11-17 October 1998.

اقتصاد دارند باز خواند کرد. با آن که متفکران گوناگون به توضیح و تبیین نگرش های متفاوت می پردازند، همگی به طور یکپارچه، تکامل جامعه اطلاعاتی را، حداقل برای چند قرن آینده، مرحلهٔ مرکزی فعالیت بشری پیش بینی می کنند. بنابراین در قرن بیست ویکم یک جامعه اطلاعاتی جهانی ظهور می کند که اطلاعات الکترونیکی کالای اصلی آن است؛ کالایی که در زندگی در سایه آن تداوم می یابد و شیوهٔ زندگی جدیدی را پدید می آورد. فن آوری بنیادین برای حمایت از جامعه اطلاعاتی، فن آوری اطلاعاتی (IT) است. بنابراین، فن آوری اطلاعات که تاکنون یک فن آوری امکان ساز (۱) شمرده می شد، اکنون فراتر می رود و نقشی مرکزی می یابد. پیشرفت های سریع در فن آوری اطلاعات به ساختار اساسی جامعه شکل می دهد و بر سازماندهی کار، چگونگی تولید و تجارت، چگونگی مدیریت، و چگونگی ایجاد ثروت در حال ظهور است. به لحاظ فن آوری، محصولات اطلاعاتی الکترونیکی از طریق دیسک های نوری یا محصولات شبکه مبنا قابل دسترسی اند. محصولات دیسک نوری که به محصولات دیسک فشرده مشهورند، و محصولات شبکه ای که غالبا" با نام محصولات وب از آنها یاد می شود (بر خلاف این باور که این دو رقیب یکدیگرند) مکمل همدیگر در توزیع اطلاعات هستند. شهرت فزاینده و نیزان بسیار بالای گسترش اینترنت این تصور را به وجود آورده که شبکه ها بزودی به عنوان مکمل انتخابی در توزیع اطلاعات، جای دیسک ها نوری را خواهند گرفت. اما آمار عکس این تصور را نشان می دهد.

حرکت به سوی جامعهٔ اطلاعاتی الکترونیکی، کتابخانه ها را به سمت خودکارسازی، تأسیس شبکه های دیسک فشرده، و ایجاد محیط کتابخانه ها و رقومی کردن آن اقدامی لازم برای دسترس پذیر کردن منابع اطلاعاتی در یک شاهراه اطلاعاتی جهان همچون اینترنت است. تعدادی از کتابخانه ها دور و نزدیک دنیا دستیابی به اطلاعات از طریق اینترنت را آغاز کرده اند و توجه خود را بیش تر و بیش تر به اطلاعات شبکه مبنا معطوف می کنند. با توجه به این روند، نیاز به کتابخانه رقومی کاملا" آشکار شده است. در نتیجه، کشف دانش در کتابخانه های رقومی به موضوعی مهم بدل می شود. این مقاله بحث درباره موضوعات مرتبط با کشف دانش در کتابخانه های رقومی می پردازد. فرآیند کنونی کشف دانش در داده پایگاه ها (KDD)(۲) که استفاده گسترده ای از فنون استخراج داده ها می کند، عمدتا" با داده های سازمان یافته و مرتبط سروکار دارد. کتابخانه های رقومی که اساسا" انبارهای دانش به شکل رقومی آن هستند عمدتا" مقادیر بسیاری از اطلاعات نامرتبط و سازمان نیافته را نیز در بر می گیرند. این مقاله به بحث درباره نارسایی های فنون کنونی استخراج داده ها در کشف دانش در کتابخانه های رقومی (KDDL)(۳) می پردارد و استخراج دانش را به مثابه رویکردی در کاربردهایی از این دست پیشنهاد می کند. بخش دوم مقاله در باره فرآیندگذار، که کتابخانه دستخوش آن هستند، بحث می کند. بخش سوم، فنون کنونی استخراج داده ها را مورد بحث قرار می دهد و نقاط ضعف را در برخورد با انبارهای دانش بیان می کند. پس از بحث درباره مفهوم استخراج دانش در بخش چهارم، در بخش پنجم نتیجه گیری ارائه می شود.

دگرگونی کتابخانه

انواع گوناگون کتابخانه ها- دانشگاهی، ملی، عمومی، یا انتفاعی- در سراسر جهان دشتخوش یک فرآیند دگرگونی هستند و نیروی محرکه این فرآیند،در اصل این واقعیت است که جامعه به طور کلی در حال حرکت یخ سوی شیوه نوینی از زندگی است که در آن، اطلاعات الکترونیکی کالای محوری است. کتابخانه های سنتی برای همگان با این تحول، د رحال خودکارشدن هستند و کتابخانه های خودکار به کتابخانه های الکترونیک تبدیل می شوند و داشتن منابع دیسک فشرده و شبکه های محلی با خدمتگرهای دیسک فشرده از مشخصات این نوع کتابخانه ها است. کتابخانه ها ی دارای منابع دیسک فشرده و محیط شبکه محلی در حرکت به سوی رقومی کردن کل مجموعه خود هستند تا بتوان از طریق شبکه به کل منابع کتابخانه دست یافت؛ بدین ترتیب کتابخانه های رقومی به وجود می آیند. این کتابخانه ها که دارای منابع محدود اما تخصص فن آوری مناسب هستند توجه خود را به دستیابی شبکه ای به منابع موجود در دیگر نقاط جهان معطوف می کنند و نتیجتا" مفهوم کتابخانه های مجازی را پدید می آورند. تکامل و پیشرفت انواع گوناگون کتابخانه ها و ویژگی های آنها در جدول ۱ نشان داده شده اند.

جدول ۱. تحول کتابخانه ها

شماره	نوع کتابخانه	ویژگی
۱	کتابخانهٔ سنتی	موجودی به صورت چاپی، که هیچ کار رایانه ای روی آن نشده
۲	کتابخانهٔ خودکار	خودکار کردن فعالیت های کتابخانه، فهرست نویسی، امانت، فراهم آوری، و . . . رایانه ای؛ موجودی عمدتا" به شکل چاپی؛ تعداد محدودی منابع الکترونیکی .
۳	کتابخانهٔ الکترونیکی	فعالیت های کاملا" خودکار؛ شبکه دیسک فشرده؛ منابع به دو شکل الکترونیکی و متعارف
۴	کتابخانهٔ رقومی	کاملا" خودکار؛ تمام منابع به شکل رقومی؛ شبکهٔ محلی فیبرهای نوری با سرعت بالا و دستیابی به شبکه های گسترده
۵	کتابخانهٔ مجازی	کتابخانهٔ بودن دیوار؛ تأمین دستیابی به منابع؛ کتابخانهٔ بدون منابع

کتابخانه سنتی

دانش در طی قرن ها گرد آوری، ثبت، سازماندهی، و در گنجینه های گوناگون ذخیره شده است. عام ترین این گجینه ها کتابخانه های سنتی هستند. اصطلاح کتابخانه سنتی برای اشاره به نظام قدیمی کتابخانه هایی به کاردر آنها موادی مثل کتاب ها و مجلات به شکل چاپی در قفسه ها نگهداری می شوند و کتابدارن در بازیابی این مطالب به مراجعه کنندگان کمک می کنند. در این نوع کتابخانه ها ، چندین نوع رسانه متفاوت و مستقل برای ذخیره اطلاعات و دانش مورد استفاده قرار می گیرند. برای مثال رسانه های کاغذی، ریز فیلم، و ریز برگه به لحاظ فیزیکی نمایانگر فن آوری های ذخیره متفاوتی هستند. بتابراین، کتابخانه های سنتی

کتاب های متعارف را در کنار اطلاعاتی که بر روی نوارهای صوتی و تصویری، ریزفیلم ها، دیسک های تصویری، سازمانی و مدیریت د راین کتابخانه ها اساسا" دستی است. فرآیند های بازیابی اطلاعات که در این کتابخانه ها به کار گرفته می شود نیز برمبنای استفاده از نمایه های کارتی، ریزفیلم، نمایه های ریزبرگه ای و مانند آن ها و ماهیتا" دستی هستند. با افزایش مداوم مجموعه ها و با وجود نیروی انسانی فاقد آموزش مناسب، سازماندهی و مدیریت کتابخانه های سنتی در طی یک دورهٔ زمانی و خامت قابل ملاحظه ای را به خود پذیرفته است. دستیابی استفاده کننده به اطلاعات به فرآیندی ناگوار تبدیل شده و استفاده کننده زمانی طولانی را برای کاوش و مکانیابی اطلاعاتی که به دنبال آن است صرف می کند. از نظر تاریخی، کتابخانه ها ، برای انجام تعهدات اصلی خود فن آوری های نوینی اتخاذ کرده اند. کتابخانه های سنتی به عنوان قسمتی از این فرآیند و برای بهبود کارآیی در عملیات خود، اقدام به بهره گیری از فن آوری رایانه ای کرده اند. بنابراین کتابخانه هایی که عمدتا" دارای مواد چاپی اند در حرکت به سوی خودکار کردن فعالیت های خود از قبیل امانت و ایجاد فهرست های ماشین خوان هستند.

کتابخانه های خودکار

با گذشت زمان، انجام دستی عملیات کتابخانه ای کارآیی خود را از دست داد. افزایش مداوم تعداد مراجعات و نیز انفجار اطلاعات، بهره گیری از فن آوری رایانه ای را برای خودکار کردن و افزایش کارآیی عملیات در پی داشته است. د رکتابخانه های خودکار، کارکردهای کتابخانه ای نظیر اذاره کتابخانه، فراهم آوری منابع، امانت، کنترل نشریات، فهرستنویسی با استفاده از فن آوری رایانه ای انجام می شود. داده های کتابشناختی تک نگاشت ها، مجموعه مقالات کنفرانس ها ، گزارش ها و . . . با آماده کردن کاربرگه های گوناگون برای اسناد مختلف یا با اسکن کردن، سریعا" وارد رایانه می شوند. با این عملیات، فهرست [موجودی] کتابخانه، ماشین خوان می شود و مکانیابی اسناد را می توان بسیار بالا برد، بلکه بازیابی اطلاعات را نیز می توان از طریق نقاط گوناگون دستیابی در فهرست راهنما همچون نویسنده، عنوان، کلید واژه ها، و . . . یا با انتخاب از فهرست راهنما انجام داد . در کتابخانه خودکار فعالیت های میز امانت نظیر صدور (امانت گیری یا امانت دهی)، برگشت، و روزو کردن به وسیله رایانه انجام می شود. در نتیجه در این کتابخانه ها به انجام کاوشی ملال آور در برگه دان برای پی بردن به این که آیا سندی خارج شده یا خیر،یا اینکه برای افراد دیگری روزو شده یا خیر، نیازی نیست. علاوه بر این سازماندهی، و مدیریت کتابخانه را می توان با مطالعه آمار گردش کتاب بهبود بخشید. همچنین کنترل پیامدها، فرآهم آوری پیایند ها و بازبینی آنها، در خواست پیایند ها براساس پارامترهای کاوش، و مدیریت صحافی با استفاده از سیستم های رایانه ای انجام می شود. در کتابخانه های خودکار انجام مشاغل اجرایی و نیز سایر وظایف مانند تخصیص شماره UDC

به سند نیز با استفاده از رایانه امکانپذیر است. این کارکردها را می توان با پیکربندی رایانه ای تک کاربره، چند کاربره،یا در شبکه محلی نیز انجام داد.

کتابخانه های الکترونیکی

عبارت کتابخانه الکترونیکی به طور ضمنی به مفهوم آن است که فرآیندهای اصلی کتابخانه اساسا" دارای ماهیت الکترونیکی شوند، بدیهی است که مهم ترین شیوه تحقیق این امر، استفادۀ گستردۀ از رایانه برای دسترس پذیر کردن خدماتی چون نمایه درونخطی ، امکانات کاوش و بازیابی متن کامل، بایگانی خودکار سوابق، و تصمیم سازی مبتنی بر رایانه است. یک شاخص مهم کتابخانۀ الکترونیکی، حرکت آگاهانه به سوی استفادۀ گسترده از رسانه های الکترونیکی برای ذخیره، بازیابی، و تحول اطلاعات است. این به معنای آن است که کتابخانه های الکترونیکی اطلاعات بیش تر و بیش تری را به شکل الکترونیکی یعنی به شکل دیسک های فشرده یا دستیابی از طریق اینترنت فراهم آورند. شبکه کردن دیسک های فشرده از دیژگی های عام کتابخانه های الکترونیکی است. اگرچه در این کتابخانه ها از رسانه های الکترونیکی استفاده گسترده ای می شود، کتابهای متعارف نیز در کنار انتشارات الکترونیکی ارائه می کنند؛ در عین حال برخی از در خواست های روزمره ای که کتابداران به آنها رسیدگی می کنند خودکار می شود و به وسیلۀ رایانه مرتفع می گردد. کتابخانه های الکترونیکی همچنین ممکن است از سیستم های خبره ای استفاده کنند که آنان را قادر می سازد با بهره گیری از مجموعه قوانین مربوطه ، از تخصص بشری در حوزه های خاصی تقلید کنند. بازیابی اطلاعات در کتابخانه الکترونیکی معمولا" از طریق شبکۀ محلی انجام می شود. این شبکه، شبکه ای ارتباطی است که درمنطقۀ جغرافیایی محدودی فعالیت می کند و قابلیت اتصال بسیاری از تجهیزات جنبی مستقل مثل چاپگر، دیسک ران(۴)، مدم، و دیسک های فشرده را فراهم می آورد. استفاده از شبکۀ محلی فیبرنوری، انتقال اطلاعات با سرعت های بالاتر را ممکن می سازدو و به فعالیت قابل اطمینان تری منجر می شود. رسانه های شبکه شدۀ دیسک فشرده دستیابی سریع به حجم عظیمی اطلاعات را به طور همزمان برای کاربران متعدد فراهم می کنند و مدیریت کتابخانه می تواند کنترل متمرکز بر منابع اطلاعاتی و همچنین کنترل امنیتی را اعمال کند. کتابخانه های الکترونیکی مدتی است که فعال اند و در حال بررسی حرکت به سوی کتابخانه های رقومی هستند. در کتابخانه های رقومی کلیۀ منابع کتابخانه رقومی می شوند و در شبکه ها قابل دستیابی خواهند بود.

کتابخانه های رقومی

در کتابخانه های رقومی اطلاعات به طور الکترونیکی ذخیره و بازیابی می شوند. از آنجا که اطلاعات به لحاظ وثوق، ثبات، و قانونی بودن مطالب اطمینان حاصل کرد. هزینه های دستیابی باید قابل پیش بینی،قابل کنترل، در حد معقول و قابل قبول باشد. کتابخانه های رقومی باید

امکانات پیچیده ای را برای انجام ناوبری [اطلاعاتی] منطقه ای و جهانی فراهم کنند تا از طریق آن انجام کاوش کارآمد و اجتناب از دوباره کاری مسیر گردد. کتابخانه های رقومی علاوه بر تأمین دشتیابی به اطلاعات، باید به استفاده کنندگان مجاز امکان اضافه و تغییر ذخیره شده را بدهند. بنابراین دادن اطمینان به دارندگان حق تکثیر (کپی رایت) دربارهٔ استفاده صحیح از اطلاعات و جلوگیری از نقض قانون به وسیلهٔ کارکنان و استفاده کنندگان نیز از ابعادی است که باید در عرصه رقومی زیرپوشش قرار گیرند. به منظور دستیابی به اطلاعات رقومی لازم است که یا از ایستگاه های خوانندهٔ(۵) چند رسانه ای یک منظوره، یا از نوع دیگری از سیتم های رایانه ای استفاده کرد. این تجهیزات را می توان در محدودهٔ تالار مطالعهٔ عمومی یک منطقه ، یا در اتاق های انفرادی که برای مطالعه شخصی مورد استفاده قرار می گیرند تعبیه کرد. دستیابی به اطلاعات از ران دور و به وسیلهٔ مدم و تلفن یا شبکه های ارتباطی رایانه ای نیز امکانپذیر است. بنابراین دستیابی به کتابخانه رقومی در مزرهای مکان یا زمان محدود نمی شود و دستیابی به آن از هر جا و در هر زمانی میسر است. مطالب کتابخانه می تواند به هر یک از اشکال داده ای باشد، اگرچه شکل اصلی اتنشارات موجود د راین نوع کتابخانه، کتاب ها و مجله های الکترونیکی هستند. واضح است که کتابخانه های رقومی آینده باید در بستر اقتصادی، اجتماعی و حقوقی بسیار گسترده تری به فعالیت بپردازند. در چارچوب لازم برای کتابخانه رقومی باید این عوامل را در نظر گرفت تازمینه لازم برای ذخیرهٔ تعداد اقلام بسیار گسترده ای را به شکل رقومی دراختیار دارند و لازم است برای ذخیرهٔ سازمان یافته داده های خود مفاهیم انبارش داده ها را محقق سازند. کتابخانه های رقومی، اساسا" تالارهای بزرگ دانش الکترونیکی هستند؛ از این رو است که کشف دانش در کتابخانه های رقومی اهمیت می یابد .

کشف دانش و استخراج داده ها

کشف دانش در داده پایگاه ها (KDD) حوزه ای در شرف تکوین است که فنونی را از یادگیری ماشینی، شناسایی الگوها، آمار، داده پایگاه ها، و تجسم بخشی(۶) گرفته تا مفاهیم گزینش خودکار، روابط متقابل مفهومی، و الگوهای مورد علاقه را از داده پایگاه های بزرگ با هم ترکیب می کند فنون ،KDD به ما امکان می دهد که دانش جدید را در داده پایگا هایی که در آن ها ابعاد، پیچیدگی یا مقدار داده ها زیاد است- چنان زیاد که تنها با مشاهدهٔ مستقیم نمی توان آن ها را یافت- بیابیم . امروزه با افزایش مداوم مقدار داده های گرد آوری شده در حوزه های گوناگون و در نتیجه دشواری تحلیل دستی داده ها، قابلیت های فنون KDD اهمیت بیش از اندازه می یابند. در سه سطح می توان به کشف دانش اقدام کر: دانش همگانی، دانش در سطح ابتدایی و دانش در سطح پیشرفته .

فنون استخراج داده ها ابراز اساسی واکشی الگوها از داده ها است. استخراج داده ها شاخه ای از دانش است که به کشف دانش پنهان، الگوی پیش بینی نشده، و قواعد جدید در داده پایگاه

های بزرگ الکترونیکی می پردازد. سپس از دانش واکشی شده در کارهایی از نوع پیش بینی و رده بندی، تخلیص مطالب داده پایگاه ها، یا توضیح پدیده های مورد مشاهده استفاده می شود. فرآیند استفاده از این ابرازها مشتمل بر پیش پردازش، گزینش و تغییر شکل داده ها، و تفسیر الگوها به "دانش "، فرآیند KDD نامیده می شود. فنون استخراج داده ها عمدتا" از حوزه های مربوطه در شبکه های عصبی، آمار،رده بندی الگوها، و یادگیری ماشینی به دست می آید. رویکرد طبیعی به کشف دانش در کتابخانه های رقومی به کار بستن فنون موجود استخراج داده ها است. از این فنون به طور موفقیت آمیزی برای کاربردهای متفاوتی ا زقبیل کنترل کیفیت، تشخیص طبی و پیش بینی مخابرات، کشف تقلب در کارت اعتباری،کشف سوء استفاده و نقض امنیت در رایانه ها، بازاریابی، سرمایه گذاری مالی، و آسان سازی استفاده از شبکه جهانی وب از طریق پیش بینی سایت های سودمند براساس رفتار گذشتهٔ استفاده کنندگان استفاده می شود. ویژگی این کاربردها، داده های مرتبط و سازمان یافته در مقیاس وسیع است. برعکس، تالارهای دانش در کتابخانه های رقومی مقادیر گسترده ای از داده های سازمان نیافته در حوزه های متنوع موضوعی را در برمی گیرند. در نتیجه، فنون حاضر در استخراج داده ها فاقد مطلوبیت لازم برای KDDL است و جستجو برای سافتن فنون جدید ضرورت دارد.

فنون استخراج داده ها

فنون استخراج داده ها شامل قواعد تداعی، قواعد رده بندی، قواعد ممیزه، خوشه بندی، و تحلیل رشد و انحراف می شود. د روضعیت خاص میتوان یک فن یا ترکیبی از چند فن را برای کشف دانش اعمال کرد. میزان سودمندی فنون استخراج داده ها متفاوت است. برخی از این فنون در کشف دانش ابتدایی سودمند هستند؛ حال آن که دیگر فنون از قابلیت لازم برای کشف دانش در ژرفای بیش تر برخوردارند. بسیاری از فنون تجربی استخراج داده ها برای کاربر روی داده های مقیم در حافظه طراحی شده اند و برخی از فنون برای عمل بر روی داده پایگاهها ارتقاء یافته اند.فنون استخراج داده ها را می توان به پنج طبقهٔ بزرگ تقسیم کرد که به طور خلاصه در بند های زیر توصیف می شوند.

الگوسازی پیش بینانه

در الگوسازی پیش بینانه، هدف پیش بینی مندرجات برخی فیلدها در یک داده پایگاه برمبنای مطالب دیگر فیلدها است. اگر فیلدی که دربارهٔ آن پیش بینی می شود فیلد عددی باشد، با پیش بینی به عنوان یک مسئلهٔ رگرسیونی برخورد می شود و اگر فیلدی که پیش بینی می شود مقوله ای مثلا" متوسط، بلند، بد، خوب باشد، در آن صورت این فیلد در مقام مسئله رده ای مورد بررسی قرار می گیرد. الگوریتم های رده ای و رگرسیونی بسیار متنوعی وجود دارند. دروندادهای یک مسئله در پیش بینی شامل مطالب سایر فیلدها، داده های

آموزشی، و آگاهی قبلی از مسئله می شود داده های آموزشی داده هایی هستند که به وسیلهٔ مشاهده کنندگان گردآوری می شوند و در بارهٔ ارزشی هستند که فیلدهای هدف برای درونداد فیلدهای مورد نظر دارند. آگاهی قبلی به وسیلهٔ مجموعه ای از فرضیه ها به طور کلی مورد بحث قرار می گیرد.

خوشه بندی

خوشه بندی که به قطعه بندی نیز مشهور است، فیلدهایی را که باید پیش بینی شود مشخص نمی کند بلکه تقسیم داده ها به موضوعاتی راکه شبیه به یکدیگر هستند مدنظر خود قرار می دهد. از آنجا که تعداد خوشه های دلخواه از پیش تعیین شده نیست، الگوریتم های خوشه ای نوعاً از یک کاوش دو مرحله ای بهره می گیرد: یک مدار بیرونی برای رسیدن به تعداد معینی خوشه، و یک مدار درونی برای رسیدن به مناسب ترین و بهترین اندازه برای آن تعداد معین خوشه ها.

تخلیص داده ها

گاهی اوقات،هدف فقط دستیابی به الگوهایی موجز است که زیر مجموعه های داده ها را توصیف می کنند. نوع شیوه وجود دارد که نمایانگر برش افقی (موارد) یا عمودی(فیلدها) در داده ها است. در برش افقی، خلاصهٔ زیر مجموعه ها تولید می شود، مثل تولید آمار کارآمد، یا شرایط منطقی که برای زیر مجموعه ها مصداق دارد. در حالت دوم روابط بین فیلدها پیش بینی می شود. تفاوت این نوع از شیوه ها با نوع نخست در آن است که هدف در اینجا، یافتن روابط بین فیلدها ایت تا پیش بینی یک فیلد مشخص (رده بندی) یا گروه بندی موارد(خوشه کردن) . یک شیوهٔ رایج، قواعد تداعی نامیده می شود. تداعی ها قواعدی هستند حاکی از این که برخی ترکیبات ارزش ها(ارقام) با ترکیبات دیگری از ارزش ها(ارقام)، یا تواتر و قطعیت معین، پدیدار می شوند.

الگوسازی وابستگی

شناخت داده ها غالباً با استنتاج نوعی از ساختار علت و معمولی در داده ها به دست می آید. الگوهای علیت یا بر مبنای احتمالات است، همچون استنتاج یک گزاره دربارهٔ توزیع احتمالات حاکم بر داده ها، یا جبری است، مثل استنتاج وابستگی های کارکردی بین فیلد ها در داده ها.

تشخیص انحرف

این روش دقیقاً بر عکس قطعه بندی داده ها، و هدف آن به طور اخص عبارت است از تشخیص نقاط پرت و نامربوط در مجموعهٔ خاصی از داده ها و توضیح این که آیا وجود این نقاط به خاطر اختلالات(۷) است، به خاطر ناخالصی های دیگری است که در داده ها رخ می نماید، یا به خاطر دلایل علت و معلولی است. این روش که معمولاً در تلفیق با قطعه بندی

انجام می شود،غالبا" منبع انجام کشف حقیقی است، چرا که وجود نقاط پرت و نامربوط، بیانگر انحراف از یک هنجار و انتظار از پیش شناخته شده می باشد. این روش ها ناضر بر اطلاعات متوالی است، چه این توالی برمبنای نظم زمانی باشد یا بر مبنای نظم دیگری. ویژگی بارز این نوع روش ها آن است که ترتیب مشاهده اهمیت دارد و باید به آن توجه کرد.

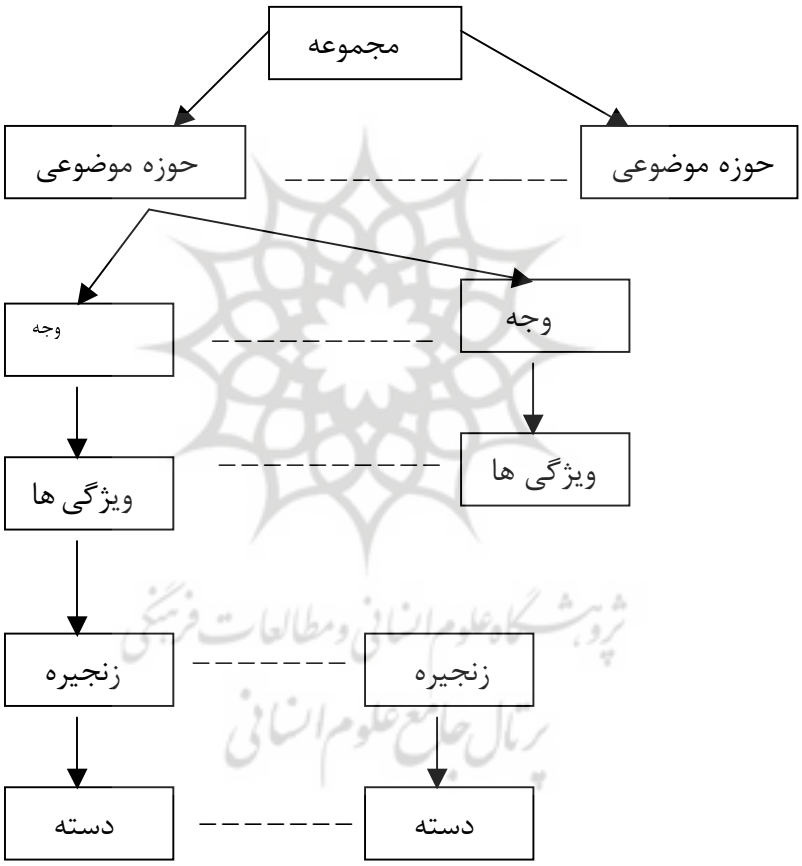
استخراج داده ها و KDDL

از توصیف فنون گوناگون استخراج داده ها می توان دریافت که این فنون عمدتا" برای پرداختن به داده های مرتبط و سازمان یافتۀ موجود در داده پایگاه ها طراحی شده اند. مجموعۀ یک کتابخانه رقویم می تواند سازمان یافته یا سازمان نیافته، متنی یا عددی، تصاویر اسکن شده، نمودار، و ضبط صوتی و تصویری مشتمل بر موضوعات متنوع باشد. در واقع کتابخانه های رقومی تالارهای بزرگ دانش الکترونیکی هستند. توانایی های فنون کنونی استخراج داده ها برای کار در این تالارهای بزرگ دانش الکترونیکی بسیار محدود است. با توجه به این موضوع، لازم است به دنبال فنونی گشت که در کتابخانه های رقومی کارآمد باشند.

فرآیند KDDL

در بافت کتابخانه های رقومی، کاربرد فنون سازماندهی دانش که در کتابخانه های سنتی به کار می رود بسیار نویدبخش به نظر می رسد. اجازه دهید نگاهی به کار آماده سازی یک سند برای استفاده در کتابخانه بیندازیم. یک متخصص موضوعی یا کتابدار، تمام یا قسمت هایی از سند را می خواند و براساس شیوۀ رده بندی مورد استفاده در کتابخانه، شماره دره ای را که باید به این سند داده شود معین می کند. روش های رده بندی متنوعی نظیر رده بندی کولن (CC)، رده بندی دهدهی جهانی (UDC)، و رده بندی دهدهی دیویی (DDC) متداول است. شماره ردۀ اختصاص یافته برای کارکنان حرفه ای کتابخانه به طور کامل و برای مراجعه کنندگان تا اندازه ای آشنا است و برای چیدن سند د رقفسه یا بازیابی آن به کار برده می شود. اکثر روش های رده بندی امکان رده بندی عمقی را می دهد. عمقی که یک سند در آن عمق رده بندی می شود، دقت دانش مندرج در آن سند را تعیین می کند. بنابراین روش های مورد استفاده در کتابخانه های سنتی دارای ساختار طبیعی هستند که امکان کشف دانش در سطوح متفاوت را فراهم می کنند. رویکردی که کتابخانه های سنتی در سازماندهی دانش دنبال می کنند در نمودار ۱ نشان داده شده. در این رویکرد مجموعۀ کاملی از اجزا به یک سری حوزه های موضوعی تقسیم می شود. بنابراین هر حوزه موضوعی یک گروه بندی کلان است، و در قالب توالی های معنادار مرتب می شود. سپس هر حوزۀ موضوعی به یک مجموعه از وجوه [(چهریزه ها)] تقسیم می گردد. بنابراین هر وجه [(چهریزه)] یک گروه بندی خاص است. وجوه [(چهریزه های)] موجود در هر حوزه موضوعی در قالب توالی های معنادار مرتب می شود. سپس هر وجه [(چهریزه)] در یک قالب سلسله مراتبی سازمان می یابد و مرحله به مرحله با به کار گرفتن مجموعه ای از ویژگی ها، به زیر بخش هایی تقسیم می گردد. این ویژگی ها با پیروی از توالی منظم ، مرتب شده اند. هر یک از توالی های اصطلاحات که به

وسیلهٔ زیربخش های بعدی ایجاد شود یک زنجیره است. هر یک از سطوح زیر بخش، گروهی از اصطلاحات را در پی خواهد داشت که یک دسته را تشکیل می دهند. اصطلاحات در هر دسته در قالب توالی های معنا داری مرتب می شود. ممکن است قواعدی را تدوین کرد که برای ترکیب اصطلاحات برگرفته از یک دستهٔ واحد، و از حوزه های گوناگون مورد استفاده قرار می گیرند. این قواعد ممکن است مستلزم استفاده از نقش نماها یا عملکردهای ارتباطی خاصی باشد. هر حوزه، وجه و اصطلاح برای تثبیت موقعیت آن در کل سیستم کدبندی می شود تا ترکیب خالی از ابهام آن تا با سایر کدها تسهیل گردد. می توان برای نشان دادن کد هر یک از اصطلاحات، وجوه، و حوزه ها، نمایه الفبایی آن ها را فراهم کرد. در دنباله فعالیت هایی که در کتابخانه های سنتی انجام می شود، KDDL را می توان در یک فرآیند دو مرحله ای تصویر کرد: ۱- سازماندهی دانش، ۲- کشف دانش.



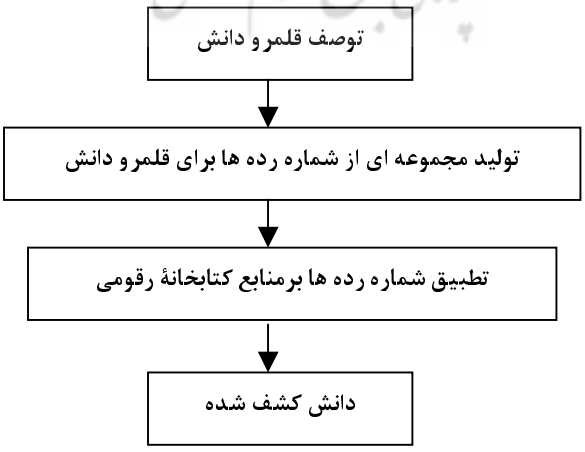
نمودار ۱. شمای سازماندهی دانش، مورد استفاده در کتابخانه های سنتی

مرحلهٔ پیشنهادی سازماندهی دانش در کتابخانه های رقومی درنمودار ۲ نشان داده شده. هرگاه قرار باشد که منبعی به یک کتابخانهٔ رقومی اضافه شود، مجموعه ای از شماره رده ها به منبع تخصیص می یابد که کاملاً محتوای دانشی منبع را توصیف می کند. در یک کتابخانهٔ رقومی ، استفاده کننده به خاطر دستیابی رایانه ای، می تواند به کل یک سند یا بخش هایی از آن دست یابد. بنابراین به فصل ها، بخش ها یا حتی بندهای یک سند می توان شماره رده های

مناسبی تخصیص داد تا بتوان دانش ویژه ای را بازیابی کرد. برعکس کتابخانه سنتی که در آن به هر سند یک یا فقط چند شماره رده اختصاص می یابد، درکتابخانه رقومی بر هر سند می تواند صدها شماره رده داد. در نتیجه در یک کتابخانهٔ رقومی هر سند را می توان بایک بردار مندی توصیف کرد که شامل شماره رده هایی می شود که ویژگی های مطالب دانشی سند را بیان می کنند. بردارهای سندی معمولا" با ماتریس های پراکندهٔ درون رایانه نمایش داده می شوند. اسناد کتابخانه های رقومی را می توان کاملا" با رایانه خواند و پردازش کرد، و می توان این کار را در فرآیند [غیر ماشینی] با استفاده از نیروی انسانی نیز انجام داد.

نمودار ۲. سزماندهی دانش در کتابخانه های رقومی

فرآیند کشف دانش در کتابخانه های رقومی در نمودار ۳ نشان داده شده. استفاده کننده ای که در جستجوی دانش است در خواست های خود را از طریق دروندادهای قلمرو دانش توصیف میکند. قلمرو دانش استفاده کننده بر روی مجموعه ای از شماره رده ها، با فرآیندی شبیه به فرآیند تصویرشده در نمودار ۲ باز نمایانده می شود. سپس این مجموعه از شماره رده ها با شماره رده هایی تطبیق داده می شود که منابع کتابخانه رقومی را توصیف می کنند. مجموعهٔ منطبق، دانش را کشف می کند.



نمودار ۳. کشف دانش در کتابخانه های رقومی

خلاصه

این مقاله پس از بحث دربارۀ فرآیند دگرگونی که کتابخانه های سراسر دنیا دستخوش آن هستند، کشف دانش در کتابخانه های رقومی (KDDL) را یکی از حوزه های مهم تحقیق دانست که باید مجادله دنبال شود. سپس به بحث دربارۀ نقاط ضعف فنون موجود، استخراج داده ها برای KDDL پرداخت. آنگاه برمبنای مفاهیم سازماندهی دانش که به طور گسترده در کتابخانه ها به کارگرفته می شوند، گروه جدیدی از فنون را پیشنهاد کرد و فرآیند KDDL را از منظر نویسندگان تشریح نمود. فرآیند KDDL که در اینجا پیشنهاد شده یک فرآیند دو مرحله ای است؛ یعنی پیش پردازش منابع کتابخانه، و فرآیند بازیابی دانش برای مراجعه کنندگان. نویسندگان بر این عقیده هستند که فنون رده بندی و سازماندهی دانش که در کتابخانه های سنتی مورد استفاده قرار می گیرند را می توان خودکار کرد و به طور مؤثر در سازماندهی و کشف دانش در کتابخانه های رقومی به کار برد.

پی نوشت ها

1. **enabling**
2. **knowledge discovery in databases**
3. **knowledge discovery in digital libraries**
4. **disk drive**
5. **reader stations**
6. **visualization**
7. **noise**

منابع

1. **Addison- Wesley Longman: Englan, 1996.**
2.
3. **discov , spain.**
4. **J.Richardson , eds , ellis Horwood series on Interactive information Systems, 1993.**
5. **Intl. Conf. IEEE comput soc., USA.1997.**
6. **implications for scien and statistical Database Management, Aug. 1997, USA.**
7. **prof. Feigenbaum to mark the opening of Aston university's new computing suite, Aston university, Birmingham, U.K. Nov .11, 1986.**

8.
Federation for Information and Dicumentation (FID), FID 714, Netherlands, 1997.
9. rge database, proc. Of the 20th intl. Conf. On very large database , sept. 1994, chile.
10. klemettinen, M.; Mannila, H.; Toivonen, H., A data mining methodology and its application to semi
Eighth Intl. Workshop on IEEE comput. Soc. Database and expert systems applications, sept. 1997 ,france.
11. Matheus , C.J,; chan , p. k. ; piatetsky-

Vol. 5 Issue: 6, Dec. 1993.
12.
Dynamic Storehouse of digitized information, 15 th Annual conventional conf. :society for information science Jan. 1996, India.
13. McClean, S.; Scotney, B., A Structured Approach to knowledge Discovery

14.
proc. Of eight Intl. Conf, on Scientific and Statistical data base Management, June 1996, Sweden .
15. Ming-

Engineering, vol. 8 No. 6 dec. 1996.
16.
fuzzy systems symposium, Dec 1996, Taiwan.
17.

1997.
18.
Kennedy . . . [et al.] data warehouse institute series from prentice hall: USA PTR, 1998.
19.
th Intl. Conf. On very large databases, dept. 1994, chile.
20. -
congrdss seminar, Aug. 1977, Denmark.
21. an Indian
pacific seminar on R& D management, Fed. 1998, philipines.
22. tentials, Vol. 16 issue: 4, oct Nov 1997, USA.
23.
Information 93, proc. Of the 17 th intl. Online information Meeting, Olympia London, 7- 9 th Dec. 1993.