# دوره توانمندسازی کارکنان دانشگاه علوم پزشکی هرمزگان

# کاربردهای داده کاوی
# برای خدمات کتابخانه های علوم پزشکی و آموزش ابزارهای آن

مدرس: نادر عالیشان کرمی
دکتری کتابداری و اطلاع رسانی
استادیار گروه فناوری اطلاعات سلامت
دانشکده پیراپزشکی بندرعباس
دانشگاه علوم پزشکی و خدمات بهداشتی درمانی هرمزگان

# تفاوت داده (Data) و اطلاعات (Information) چیست؟

- داده  یا داده‌ی خام
- اطلاعات به صورت خام و دست نخورده
- مجموعه‌ای از اعداد و ارقام درهم و شاید بدون معنای خاص
- در نتیجه پردازش و آنالیز، داده‌ها به اطلاعات تبدیل می‌شوند
- اطلاعات و ورودی‌های خام = **داده**
- نتایج حاصل از پردازش داده‌های خام = **اطلاعات**

- به عنوان مثال نمرات دانشجویان یک کلاس، داده و نتایج حاصل از پردازش روی این نمرات، شامل میانگین، میزان افزایش یا کاهش نمرات، نمودارها و... می‌توانند به عنوان اطلاعات در نظر گرفته شوند.

https://www.hamyarit.com/blog/data-mining/

# سلسله مراتب داده تا پایگاه داده

- **بیت**: به کوچکترین واحد حافظه (Memory) بیت می‌گویند. باید بدانید عظمت دیتابیس‌ها روی همین بیت‌های کوچک بنا شده است.
- **بایت**: هر ۸ بیت، یک بایت را ایجاد می‌کنند. هر بایت هم نشان‌دهنده‌ی یک کاراکتر است.
- کاراکتر: کاراکتر یا نویسه، اولین واحد منطقی داده است که انسان می‌تواند آن را درک کند. عدد، حرف یا هر علامت قراردادی‌ای می‌تواند یک کاراکتر به حساب بیاید.
- **فیلد**: فیلد یا میدان، یک مورد از اطلاعات فرد، شی یا یک پدیده است.
- **رکورد**: رکورد یا سابقه‌ی اطلاعاتی، از ترکیب چند فیلد به‌وجود می‌آید. شرح اطلاعاتی درباره‌ی فروش محصول یا مشخصات مشتری و یا مشخصات یک کالا همگی نمونه‌هایی از یک سابقه‌ی اطلاعاتی (رکورد) به حساب می‌آیند.
- **فایل**: مجموعه رکوردهای مرتبط با هم یک فایل یا پرونده را شکل می‌دهند.
- **جدول**: جدول‌ها مهم‌ترین سطح یک دیتابیس محسوب می‌شوند. هر جدول سطر و ستون‌هایی دارد که در داده‌ها در آن ذخیره‌سازی، دسته‌بندی و سامان‌دهی می‌شوند.
- **پایگاه اطلاعات**: در آخر، مجموعه‌ی جدول‌ها در کنار هم بانک اطلاعاتی یا دیتابیس را شکل می‌دهند.

# دیتا بیس یا پایگاه داده چیست؟

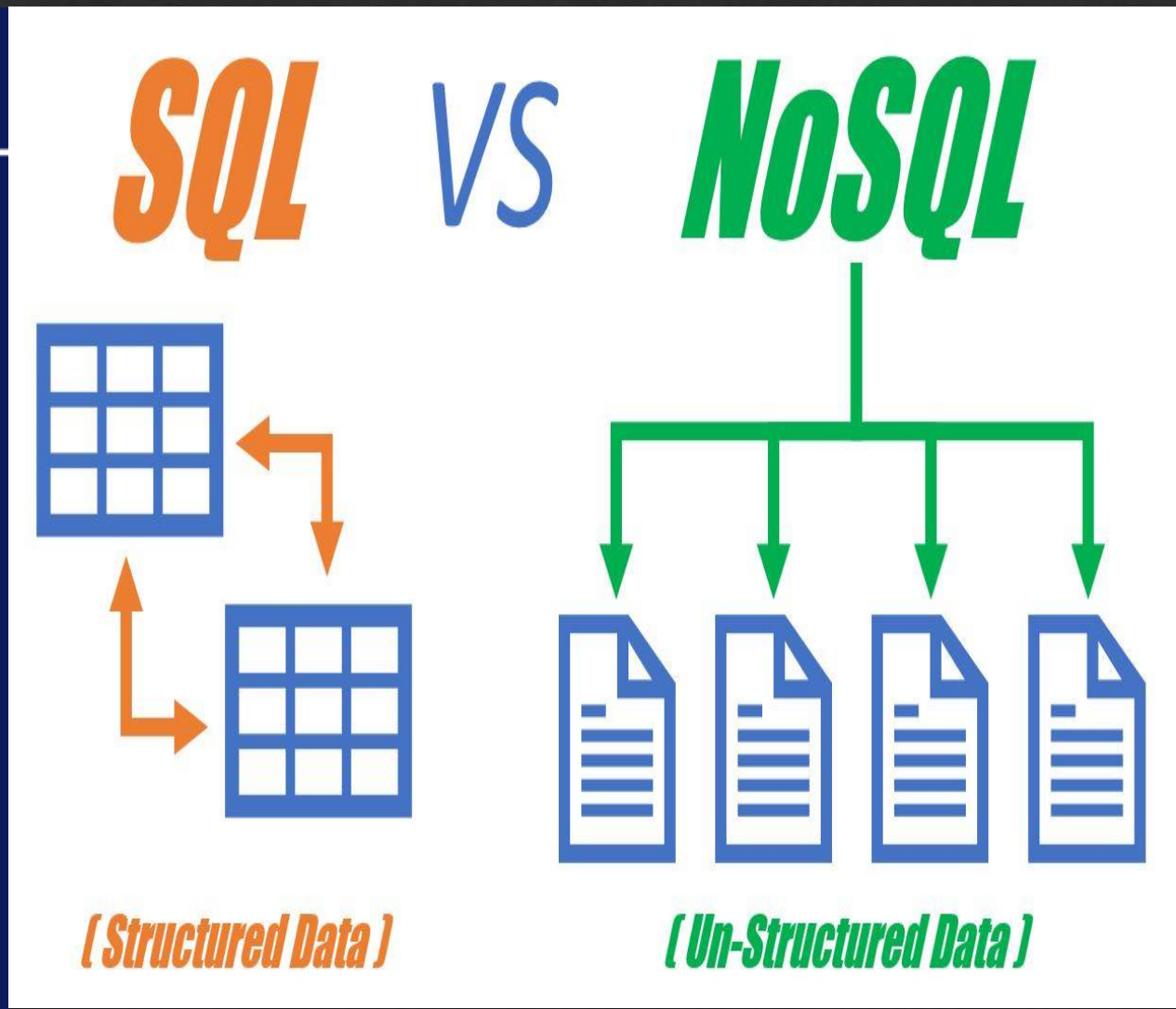
- «پایگاه داده مجموعه‌ای از داده‌های ذخیره‌شده و ثابت است که به‌صورت یک سیستم، بر پایه‌ی یک ساختار مشخص و به شکل صوری (با حداقل افزونگی) تعریف شده است. یک سیستم کنترل متمرکز این مجموعه را مدیریت می‌کند و ممکن است یک یا چند کاربر به‌طور همزمان از این مجموعه‌ی اطلاعاتی استفاده کنند.»
- «دیتابیس یا پایگاه داده یا همان بانک اطلاعات، مجموعه‌ای از داده‌هاست که در جدول‌هایی با ساختار منظم دسته‌بندی شده‌اند. این جدول‌ها همگی با هم ارتباط دارند، هرچند می‌توانند مستقل از یکدیگر هم عمل کنند.»
- داده (Data): داده‌ها نمودی از مفاهیم، معلومات، وقایع و پدیده‌ها هستند که از طریق مشاهده یا تحقیق به‌دست می‌آیند.
- اطلاعات(Information): اطلاعات درواقع همان مفهمومی است که بعد از پردازش به‌صورت داده ذخیره می‌شوند.
- موجودیت (Entity): موجودیت همان فرد، شی یا پدیده‌ای است که درباره‌اش اطلاعات جمع‌آوری شده است.
- صفت خاصه (Attribute): هر ویژگی‌ای که یک موجودیت را از موجودیت دیگر جدا کند، یک صفت خاصه محسوب می‌شود.
- سوال: سیستم مدیریت داده دقیقاً چیست و چه کار می‌کند؟ سیستم مدیریت داده یا Database Management system (DBMS)، بین دیتابیس و مدیر دیتابیس ارتباط برقرار می‌کند. درحقیقت، DBMS (از طریق زبان SQL یا هر زبان دیگری) دستورات لازم را از مدیر دریافت و در پایگاه داده اجرا می‌کند.

# اجزای اصلی دیتابیس


- **سخت‌افزار**: سخت‌افزارها از عناصر پردازشی به حساب می‌آیند. هر بانک اطلاعاتی بسته به نیازش ممکن است از سخت‌افزارهای متفاوتی استفاده کند؛ از جمله سخت‌افزارهای ذخیره‌سازی داده، سخت‌افزارهای ارتباطی، سخت‌افزارهای جانبی و ...
- **نرم‌افزار**: نرم‌افزارها به کاربر این امکان را می‌دهند تا با دیتابیس ارتباط برقرار کند؛ درست مثل یک پل ارتباطی. سیستم‌عامل، نرم‌افزارهای ارتباطی شبکه، نرم‌افزار مدیریت دیتابیس و اپلیکیشن‌ها در این دسته قرار می‌گیرند.
- **کاربر**: کاربران افرادی هستند که به روش‌های مختلفی با دیتابیس ارتباط دارند.
  - برنامه‌نویس (DataBase Programmer): افرادی که ساختار دیتابیس را طراحی می‌کنند.
  - طراحان دیتابیس (DataBase Developer): افرادی که به کمک زبان‌های مختلف از جمله SQL دیتابیس‌ها را می‌سازند.
  - مدیر پایگاه داده (DataBase Administrator): افرادی که تخصصشان، «دانش مدیریت اطلاعات» است و دیتابیس را مدیریت می‌کنند.
  - کاربران نهایی (End Users): کسانی که از داده‌ها استفاده می‌کنند.

# SQL vs NoSQL

| SQL | NOSQL |
| --- | --- |
| Relational Database management system | Distributed Database management system |
| Vertically Scalable | Horizontally Scalable |
| Fixed or predifined Schema | Dynamic Schema |
| Not suitable for hierarchical data storage | Best suitable for hierarchical data storage |
| Can be used for complex queries | Not good for complex queries |



# دیتا مارت چیست؟

- اغلب کوچک است و بر یک موضوع یا دپارتمان خاص متمرکز است
  - پاسخگوی یک نیاز داخلی ساختار اداری
- ساختار اتصال ستاره ای
- بهینه برای نیازهای کاربران دپارتمان
- رکوردهای ذخیره شده در دیتامارتها بخوبی نمایه شده اند
- دیتامارت وابسته: تامین داده ها از منابع داده ای بسیار سازماندهی شده مثل انبار داده ها
  - از لحاظ ساختاری و معماری منطقی
  - منبع دیتامارتهای وابسته تکنولوژِی بانک اطلاعات دپارتمانی
- دیتامارتهای مستقل:
  - ثابت نیستند
  - از لحاظ معماری بسیار با هم متفاوتند ← ایجاد مشکل در هنگام یکپارچه سازی دیتامارتهای مستقل
  - با یکپارچه سازی ساده دیتامارتها یک انبار داده ایجاد نخواهد شد
- طراحی دیتامارت اساسا برای اهداف تاکتیکی طراحی شده است (هدف: تامین یک نیازتجاری فوری)

# انبار داده یا Data warehouse چیست؟

## ساختار یک انبارداده

- **وابستگی به زمان:**
- نگهداری رکوردها بر اساس یک برچسب زمانی
- ایجاد صفحات زمانی بر اساس وابستگی زمانی به منظور تسهیل درک ترتیب زمانی وقایع
- **غیر فرار بودن (Non-volatility)**
- رکوردهای داده در انبار داده ها هرگز بطور مستقیم روزآمد نمیشوند.
- برای هر تغییری در ابتدا داده های عملیاتی روزآمد میشوند و سپس بگونه ای مقتضی به انبار داده منتقل میشوند. این مساله ثبات داده ها را برای استفاده های وسیعتر تضمین میکند.
- **تمرکز موضوعی**
- داده ها از بانکهای اطلاعاتی عملیاتی بصورت گزینشی به انبار داده منتقل میشوند. این استراتژی به ایجاد یک انبار داده بر اساس یک مطلب یا موضوع خاص کمک میکند و بنابراین کاوش انبار داده ها برای پرس و جوهای موضوعی با سرعت بیشتری انجام میشود.
- **یکپارچگی**
- داده ها بگونه ای کامل سازماندهی شده اند تا با حذف موارد تکراری و چند عنوانه یکپارچگی رکوردها حفظ شود ؛ به ایجاد ارجاع های متقابل کارآمد بین رکوردها کمک نموده و ارجاع دهی را تسهیل نماید.

- انبار داده کاملا " متفاوت از دیتامارت
- پوشش کلیه موضوعات حول فعالیتهای کاری سازمان از طریق سازماندهی انبارهای داده
- انبار داده نمایانگر یک تسهیلات مرکزی است.
- داده های دیتامارت به شکل خلاصه تر و متراکم تر وجود دارند
- ذخیره سازی داده ها در یک انبار داده در یک سطح نامتراکم
- ساختار داده ها در یک انبار داده یک ساختار لزوما" هنجار شده است. بدین معنی که ساختار و محتوای داده ها در انبار داده منعکس کننده ویژگیهای دپارتمانهای عضو نیست
- داده ها در انبار داده از نظر حجم و شکل کاملا" متفاوت از داده ها در دیتامارت هستند
- دیتامارت ممکن است شامل حجم زیادی از داده های قدیمی و گذشته نگر باشد
- نمایه سازی سبک داده ها در انبار داده (به بیان دیگر در عمق کمتر).
- طراحی انبار داده برای اهداف برنامه ریزی بلندمدت و راهبردی (برخلاف سیستم عملیات که کاربرمدار است متمرکز بر اقلام است)

# انبار داده چگونه کار می کند؟

- هدف ساخت انبار داده: حمایت از "پرس و جوهای" (Queries) پشتیبان تصمیم گیری (Decision Support System)
- طراحی سازمانی و عملیاتی انبار داده در راستای پاسخگویی نیازهای اطلاعاتی روزمره یا معمولی
- لزوم به کارگیری یک سیستم کامپیوتری پیشرفته برای عملیات انبارسازی داده ه بدلیل حجم بسیار بالای پایگاه اطلاعاتی
- لزوم ایجاد یک بانک اطلاعات مجزا شامل ابرداده (مشخصه هایی نظیر نوع، فرمت، مکان و پدیدآورندگان داده های ذخیره شده در یک انبار داده ها)
- هدف بانک ابرداده: کمک به کاربران و مدیران داده ها
- پردازش مبتکرانه داده های انبار داده = تولید اطلاعاتی که در وهله اول آشکار نیستند
- کشف الگوها و رابطه ها از داده های انبار داده با انتخاب متناسب داده ها، بکار گرفتن فنون مختلف غربال کردن و تفسیر زمینه ای [Contextual interpretation]
  - ایجاد بینش نو در تصمیم گیری به واسطه کشف الگوها یا رابطه های نو در داده ها
  - داده کاوی در اصل لزوما" نیاز به سازماندهی یک انبار داده ندارد

# داده کاوی چیست؟

- "به فرایند استخراج و کشف همبستگی‌ها و الگوهای مفید از میان حجم زیادی از داده‌های خام که با استفاده از الگوریتم و سازوکارهای هوشمند انجام می‌گیرد دیتاماینینگ یا داده کاوی می‌گویند، به زبان ساده‌تر، استخراج دانش از میان مجموعه‌ای از داده‌ها را داده‌کاوی می‌نامند"
- "داده‌کاوی، به مفهوم استخراج اطلاعات نهان یا الگوها و روابط مشخص در حجم زیادی از داده‌ها در یک یا چند بانک اطلاعاتی بزرگ گفته می‌شود"
- الزامات عملکرد بهینه الگوریتم های داده کاوی
  - نیاز به یک‌سری پیش‌پردازش بر روی داده‌های اولیه
  - نیاز به یک‌سری پس‌پردازش بر روی اطلاعات خروجی

https://www.hamyarit.com/blog/data-mining/

- دو کارکرد اصلی داده کاوی
  - توصیف تحلیلی داده ها
    - مثالی از تحلیل توصیفی: تحلیل داده مربوط به مشخصه های انتخابی متغیرها؛ از گذاشته و حال، و درک الگو
  - کمک به پیش بینی
    - مثالی از توانایی پیشگویانه داده کاوی: برآورد ارزش آینده یک متغیر و طرح ریزی کردن روند
- چند گام ابتدایی برای عملی شدن داده کاوی
  I. انتخاب داده ها
  II. پاک سازی داد ها
  III. غنی سازی داده ها
  IV. کد گذاری داده ها

- در قدم اول خوب بدانید هدف کلی از داده کاوی چیست؟
- اولین ضرورت: انتخاب مجموعه داده های اصلی برای تحلیل
  - استخراج رکوردهای لازم از انبار داده ها و یا بانک اطلاعاتی عملیاتی
  - پاکسازی رکوردها از داده های آلوده به منظور اطمینان از یکدستی فرمت (شکلی) آنها
    - حذف موارد تکراری
    - کنترل سازگاری دامنه
  - غنی سازی
    - گردآوری موارد ناقص یا ناکافی داده های گردآوری شده در راستای تکمیل بانک اطلاعات اصلی
    - شناسایی منابع مناسب برای تکمیل موارد ناقص
  - تعبیه یک سیستم کدگذاری مناسب جهت انتقال داده ها به فرم ساختار-بندی شده جدید؛ متناسب برای عملیات داده کاوی

# Origin of the term ‘Data Mining’

- 1960s, statisticians: Data Fishing or Data Dredging
  - to refer to the process of analyzing data without a theoretical hypothesis
- Around 1990, database community: “Data Mining”
  - Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction
  - Coining the term "Knowledge Discovery in Databases" 1st by Gregory Piatetsky-Shapiro
    - More popular in
      - Artificial Intelligence (AI)
      - Machine Learning Community.

# یک مثال قابل لمس از داده کاوی

- تبلیغات گزینشی در گوگل
- تبلیغات گزینشی در یوتیوب
- استخراج علائق کاربران از سایت های بازدید شده آنان
- استخراج الگوها و رفتارهای جستجو کاربران از داده‌های خام و به ظاهر نامرتبط و بی اهمیت
- الگوریتم‌های پیشرفته‌ی داده کاوی گوگل و تولید درآمدهای بی نهایت تبلیغاتی

https://www.hamyarit.com/blog/data-mining/

# کاربردهای اصلی داده کاوی

- کمک به کاربران جهت حذف داده‌های نامرتبط و بلااستفاده از مجموعه‌ی داده ها
- در اختیار قرار دادن اطلاعات بسیار مفید و کاربردی
- تسریع فرایندهای تصمیم‌گیری در سازمانها

**از طریق**

- **کشف الگوی میان داده‌ها**
- **پیش‌بینی تقریبی نتایج**
- **به‌دست آوردن اطلاعات کاربردی**
- **تمرکز بر روی داده‌های بزرگ**

https://www.hamyarit.com/blog/data-mining/

# Difference between Data mining and Data processing

https://theintactone.com/2022/02/22/difference-between-data-mining-and-data-processing/

# Difference between Data mining and Data Processing

## Data Processing:

Also known as Data Warehousing is a technology that aggregates structured data from one or more sources in order to compare and analyze rather than transaction processing.

### Benefits of Data Processing:

- Consistent and quality data
- Reduce in cost
- Timely access of data
- Improved performance and productivity

https://theintactone.com/2022/02/22/difference-between-data-mining-and-data-processing/

## Data Mining:

It is the process of extracting useful information, finding patterns and correlations within large data sets to identify relationships between data. Data mining tools also allow businesses to predict customer behavior.

**Benefits of Data Mining:**

- Direct marketing
- Analyzing trends
- Fraud Detection
- Forecasting in financial markets

https://theintactone.com/2022/02/22/difference-between-data-mining-and-data-processing/

| | Data Mining | Data Processing |
|---|---|---|
| **Definition** | It is the process of extracting important pattern from large datasets. | It is the process of analysing and organizing raw data in order to determine useful information's and decisions |
| **Function** | It is used in discovering hidden patterns in raw data sets. | In this all operations are involved in examining data sets to fine conclusions. |
| **Data set** | In this data set are generally large and structured. | Dataset can be large, medium or small, Also structured, semi structured, unstructured. |
| **Visualization** | It generally does not require visualization | Surely requires Data visualization. |
| **Goal** | Prime goal is to make data useable. | It is used to make data driven decisions. |
| **Required Knowledge** | It involves the intersection of machine learning, statistics, and databases. | It requires the knowledge of computer science, statistics, mathematics, subject knowledge AI/Machine Learning. |
| **Models** | Often require mathematical and statistical models | Analytical and business intelligence models |
| **Known as** | It is also known as Knowledge discovery in databases. | Data analysis can be divided into descriptive statistics, exploratory data analysis, and confirmatory data analysis |
| **Output** | It shows the data tends and patterns. | The output is verified or discarded hypothesis |

https://theintactone.com/2022/02/22/difference-between-data-mining-and-data-processing/

# فرایند انجام Data Mining


- جمع‌آوری داده‌های مورد نیاز (داده‌های هدف)
- پردازش و پاکسازی: حذف داده‌های اضافه و ورود داده‌های مورد نیاز به سیستم
- کشف الگوی میان داده‌ها و ارزیابی آن
- اجرای الگوریتم و متدهای Data Mining بر روی داده‌ها
- نمایش اطلاعات به‌دست آمده از فرایند داده‌کاوی در قالب فرمت‌های قابل درک برای انسان مانند نمودار، تصویر، گزارش و...
- ارائه دانش حاصله از انبوه داده‌های خام به سازمان

https://www.hamyarit.com/blog/data-mining/

# مشکلات اساسی بر سر راه داده کاوی

1. **حجم بالای داده‌های موجود در ورودی**

**راهکارها**

1. استفاده از الگوریتم‌های سریع‌تر، روش‌های کاهش پیچیدگی زمانی، بهینه‌سازی، پردازش موازی
2. استفاده از روش‌هایی مانند نمونه گیری، گسسته‌سازی، کاهش ابعاد و... جهت کاهش حجم داده‌های ورودی
3. استفاده از قابلیت‌های ذخیره و بازیابی اطلاعات موجود در دیتابیس‌ها از روش‌های ارائه‌ی رابطه‌ای

1. **عدم اطمینان کامل به اطلاعات خروجی**

**راهکار**: کنترل داده های ورودی

۱. کامل نبودن داده های ورودی( یعنی در داده‌ها مشخصه‌هایی وجود دارد که مقدار معتبری برای آن‌ها درج نشده است)

۲. اطلاعات ناسازگار باشند (داده‌ها با تداخل رو به رو شده باشند) و در نتیجه مقادیر ثبت‌شده با مقادیر واقعی یکسان نباشند

برطرف شدن مشکلات بالا⬅افزایش صحت داده‌های ورودی⬅اطمینان به اطلاعات خروجی حاصل از داده کاوی

https://www.hamyarit.com/blog/data-mining/

# پلتفرم‌های مورد استفاده در فرایند داده‌کاوی

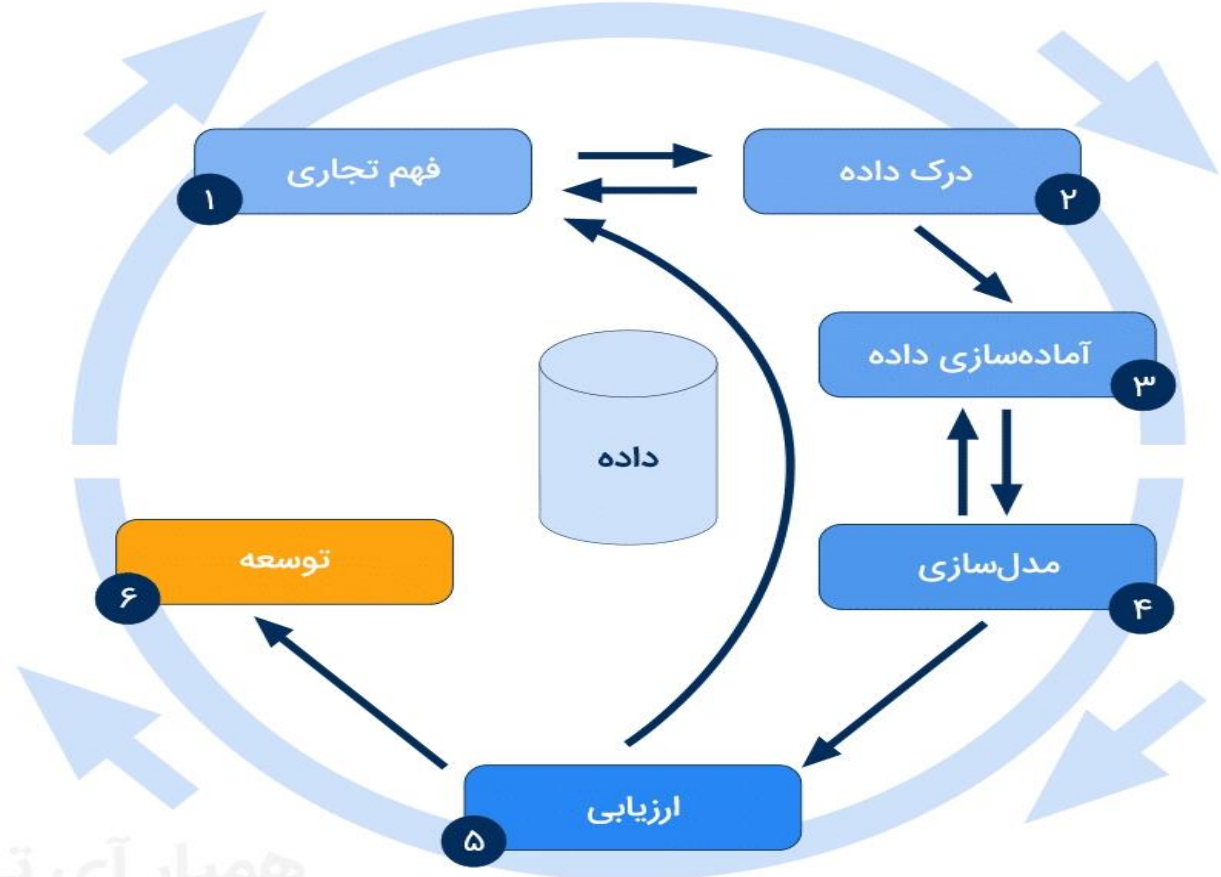
- نرم‌افزار SPSS
- نرم‌افزار Weka
- نرم‌افزار RapidMiner
- زبان برنامه‌نویسی آر (R)
- زبان برنامه‌نویسی پایتون
- زبان برنامه‌نویسی متلب

https://www.hamyarit.com/blog/data-mining/

# معرفی الگوریتم CRISP یا کریسپ در داده‌کاوی

# CRoss Industry Standard Process for Data Mining

## فرایند استاندارد داده کاوی بین صنایع




- فهم تجاری Business Understanding
- درک داده Data Understanding
- آماده‌سازی داده Data Preparation
- مدل‌سازی Modeling
- ارزیابی Evaluation
- توسعه Development

https://www.hamyarit.com/blog/data-mining/

# متدولوژی خوشه‌بندی CRISP برای فرایند داده‌کاوی

- **فهم تجاری**: شامل گردآوری موارد مورد نیاز و گفتگو با مدیران ارشد برای تعیین اهداف.
- **درک داده**: نگاه نزدیک و بررسی دسترسی به داده‌ها برای فرایند داده کاوی که شامل گردآوری، توصیف، کشف و تغییر کیفیت داده‌ها می‌شود.
- **آماده سازی داده**: این مرحله یکی از مهم ترین و همچنین زمان‌برترین بخش‌های داده‌کاوی است که شامل انتخاب، پاک‌سازی، ساختاربندی، و ادغام داده‌ها می‌شود.
- **مدل سازی**: هم‌اکنون داده‌ها آماده‌ی فرایند داده‌کاوی‌اند و نتایج راه حل‌هایی را برای مشکل تجاری مطرح شده نشان می‌دهند، تکنیک‌های انتخاب مدل‌سازی، ایجاد یک طراحی آزمون، ساخت مدل‌ها، و ارزیابی مدل این مرحله را می‌سازند.
- **ارزیابی**: در این مرحله نتایج ارزیابی شده، فرایند انجام کار بازبینی و مراحل بعدی انجام می‌شوند.
- **توسعه**: نتایج به‌دست آمده توسعه یافته و برای بهبود عملکرد سازمان به کار گرفته می‌شوند.

https://www.hamyarit.com/blog/data-mining/

# چهار نوع دانش قابل تولید در داده کاوی
# Fayyad et al. (1996)

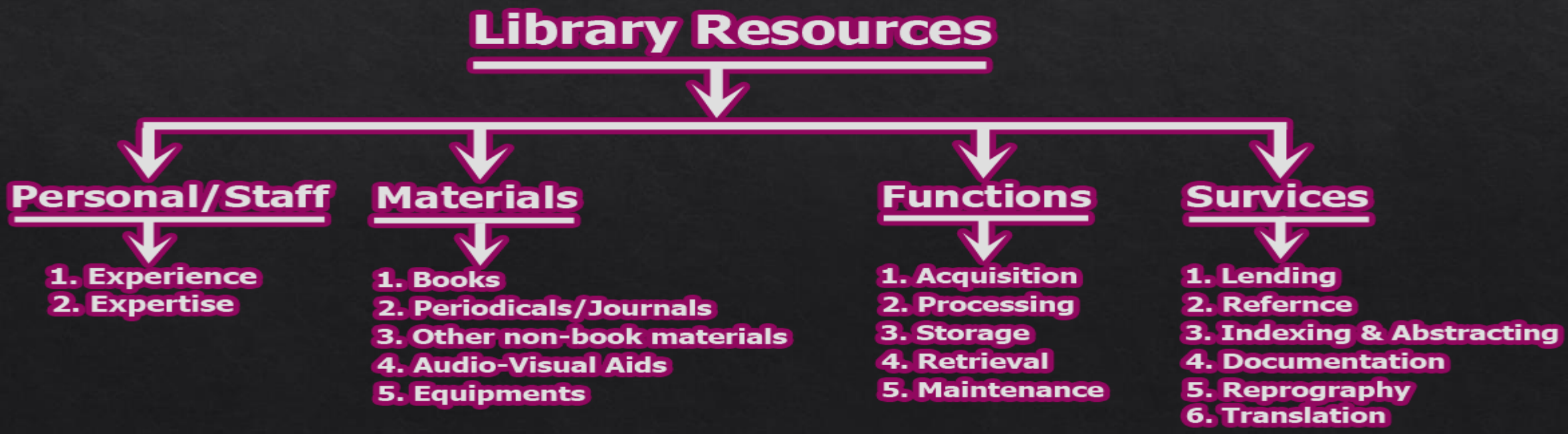
- دانش سطحی  (کاربردهای SQL)
- دانش چند وجهی (کاربردهای OALP)
- دانش نهان (تشخیص الگو و کاربردهای الگوریتم یادگیری ماشینی)
- دانش عمیق(کاربردهای الگوریتم بهینه سازی داخلی)

# کاربردهای داده کاوی در کتابخانه ها و محیط های دانشگاهی

- سرآغاز داده کاوی: حوزه تجارت
- مفید وافع شدن داده کاوی در سایر حوزه های درگیر در گردآوری حجم وسیع داده ها (حوزه های دستخوش تغییرات پویا)
  - مثل بانکداری، تجارت الکترونیک، تجارت سهام، بیمارستان و هتل
  - پتانسیل بالای استفاده از داده کاوی در حوزه آموزش
- سوالاتی که داده کاوی می تواند پاسخ دهد:
  - میزان اشتراک مورد انتظار برای نشریات بین المللی انتخاب شده برای سال آینده چقدر می تواند باشد؟
  - الگوی استفاده کلی مجلات الکترونیکی یا تحلیل درخواستهای اعضا برای میکروفیلمها طی ۵ سال گذشته چگونه است؟ (مثالهایی از کشف روندهای عمومی کتابخانه)
  - دامنه تحلیل استنادی هم میتواند با استفاده از داده کاوی گسترش داده شود.

# مدیریت و خدمات کتابخانه

- **عملیات کتابداری**
  - مدیریت مدارک
  - ارائه خدمات (انتخاب، سفارش، سازماندهی، امانت و ...)
  - امور اداره و نگهداری
- **سر و کار کارکردها/خدمات با انواع مختلف داده ها**
  - پردازش جداگانه داده ها در بخش های مختلف
  - پردازش/تحلیل ترکیبی داده ها معادل گشایش افق تازه
  - طرح خدمات جدید
  - تحول رویه ها و عملیات جاری جدول یک برخی از کاربردهای ممکن داده کاوی را که میتواند در کتابداری مفید باشد ارائه میکند.



# برخی ا زپتانسیل های کاربرد داده کاوی در کتابخانه ها

- استفاده از بانک های اطلاعاتی جهت تعیین نقاط قوت و ضعف مجموعه
- بررسی داده های گردآوری منابع جهت ایجاد رابطه بین خواننده، منابع کتابخانه و زمان مشخصی از سال در خصوص استفاده از مجموعه
- تحلیل سفارشهای پاسخ داده شده و سفارشهای دریافت شده در امانت بین کتابخانه ای
- پیش بینی روند بازگشت منابع در داده های بخش امانت
- تعیین وضعیت منابع مالی بکار گرفته شده با استفاده از داده کاوی داده های هزینه ای

فصلنامه مدیریت اطلاعات و دانش‌شناسی
سال دوم، شماره سوم. پاییز ۱۳۹۴ (۲۳– ۳۲)

Quarterly Journal of Knowledge and Information Management
Vol. 2, No. 3. Autumn 2015

# کاربردهای داده‌کاوی در علم اطلاعات و دانش‌شناسی

مهدی رحمانی[۱*]، محسن حاجی زین العابدینی[۲]
۱. دانشجوی کارشناسی ارشد، علم اطلاعات و دانش‌شناسی، دانشگاه شهید بهشتی تهران.
۲. استادیار، علم اطلاعات و دانش‌شناسی، دانشگاه شهید بهشتی تهران.
دریافت: (۱۳۹٤/۰٦/۰۳)  پذیرش: (۱۳۹٥/۰٤/۲٤)

## چکیده

هدف: داده‌کاوی یکی از راه‌های تحلیل و استفاده از اطلاعات است که با استفاده از روش‌های تخصصی آماری و منطقی به تحلیل داده‌های بزرگ‌مقیاس می‌پردازد و به موسسات در اخذ تصمیم‌های کلان کمک می‌کند. پژوهش حاضر با هدف شناسایی مهم‌ترین کاربردهای داده کاوی در حوزه علم اطلاعات و دانش‌شناسی براساس کاربردهای داده‌کاوی در سایر حوزه‌های علمی بود.

روش‌شناسی: در پژوهش حاضر جهت شناسایی کاربردهای داده‌کاوی علاوه بر مطالعات کتابخانه‌ای با تحلیل و بررسی کاربردهای داده‌کاوی در حوزه‌های بانک‌داری، مدیریت ریسک، هتل‌داری، مدیریت، تجارت، آمار و صنعت کاربردهای داده‌کاوی در حوزه علم اطلاعات و دانش‌شناسی ارائه شد.

یافته‌ها: یافته‌ها نشان داد داده‌کاوی در بخش‌های مختلفی از حوزه علم اطلاعات و دانش‌شناسی از جمله جست‌وجوی اطلاعات، فراهم‌آوری، مدیریت و خدمات کتابخانه، جمع‌آوری اطلاعات مراجعان به‌کتابخانه، حفظ وفاداری مراجعان، اخراج مراجعه‌کننده، تحلیل منابع موجود در کتابخانه (موجودی کتابخانه)، بانک اطلاعاتی، گردآوری منابع، استفاده از مجموعه و امانت بین‌کتابخانه‌ای کاربرد دارد، همچنین مشخص شد که از داده‌های بخش امانت و داده‌های هزینه‌ای جهت بهتر شدن عملکرد کتابخانه‌ها و بیشتر شدن هزینه–سودمندی کتابخانه استفاده کرد.

نتیجه‌گیری: کاربردهای داده‌کاوی در حوزه علم اطلاعات و دانش‌شناسی بیشتر مختص به بخش‌های امانت، مرجع، و مجموعه سازی است. اما می‌تواند در سایر بخش‌ها نیز از داده‌کاوی برای تحلیل داده‌های موجود در راستای دستیابی هر چه بهتر به اهداف رشته و کتابخانه‌ها استفاده کرد.

- جستجوی اطلاعات
- فراهم آوری
- مدیریت و خدمات کتابخانه
- جمع آوری اطلاعات مراجعان به کتابخانه
- حفظ وفاداری مراجعان
- اخراج مراجعه کننده
- تحلیل منابع موجود در کتابخانه (موجودی کتابخانه)
- بانکهای اطلاعاتی
- گردآوری منابع
- استفاده از مجموعه و امانت بین کتابخانه ای
- استفاده از داده های بخش امانت و داده های هزینه ای جهت بهتر شدن عملکرد کتابخانه ها و بیشتر شدن هزینه –سودمندی کتابخانه

مقاله اصیل پژوهشی

# مروری نظام‌مند برکاربرد داده‌کاوی در کتابخانه‌های دیجیتال

***شـــهناز خادمی زاده:** دانشـــیار گروه علم اطلاعات و دانش‌شـــناســـی دانشـــگاه شـــهید چمران اهواز، اهواز، ایران. (نویســـنده مســـئول)
s.khademi@scu.ac.ir
**زینب محمدی:** دانشجوی دکتری علم اطلاعات و دانش‌شناسی دانشگاه شهید چمران اهواز، اهواز، ایران

## چکیده

**زمینه و هدف:** پژوهش حاضـر با هدف شـناسـایی کاربردهای داده‌کاوی در ارائه خدمات، مجموعه‌سـازی و مدیریت کتابخانه‌های دیجیتالی صورت گرفته است.

**روش پژوهش:** پژوهش حاضـــر از نظر هدف از نوع مطالعات کاربردی و به لحاظ روش در زمره پژوهش¬های کیفی اســـت که به روش مرور نظام‌مند انجام شـده اسـت. برای این منظور مقالات از طریق جسـتجو در پایگاه¬های اطلاعاتی«اشـپرینگر»، «امرالد»، «پروکوئسـت»، «وب‌آو سـاینس»، «گوگل اسـکالر»، «سـاینس دایرکت» و «سـمنتیک اسـکالر» بدسـت آمده اسـت. مقالات بین بازهٔ زمانی ۲۰۰۰ تا ۲۰۲۱ تحلیل و از الگوی مرور سـیسـتماتیک کیچنهام و چارتر (۲۰۰۷) پیروی شـد. با توجه به معیارهای موردنظر تعداد ۱۲۹۶ مقاله بعد از پالایش اولیه اسـتخراج شـده اسـت و از بین این مقالات با بررسـی عنوان ۷۷ مقاله مرتبط با موضـوع شـناسـایی و وارد بررسـی نهایی از طریق بررسـی متن کامل شـــده اســـت که در نهایت تعداد ۲۹ مقاله برای تحلیل نهایی انتخاب شـــدند. برای تجزیه و تحلیل اطلاعات از روش کدگذاری محتوای کیفی اسـتفاده و تجزیه‌وتحلیل محتوا از سـوی دو کدگذار انجام شـد. میزان توافق ارزیابان براسـاس فرمول مایلز و هابرمن برای تحلیل‌های انجام شده ۷۸/۵ محاسبه شد.

**یافته‌ها:** بر اساس نتایج تحلیل کیفی در این پژوهش، ۷۵ مضمون پایه، ۱۳ مضمون سازمان‌دهنده و ۳ مضمون فراگیر «خدمات دیجیتال»، «مدیریت کتابخانه دیجیتال» و «مجموعه‌سازی دیجیتال» شناسایی شده است که در مجموع کاربرد داده‌کاوی در کتابخانه‌های دیجیتال را به تصویر کشیده است.

**نتیجه‌گیری:** با اســـتفاده از تکنیک‌های داده‌کاوی در کتابخانه‌های دیجیتال می‌توان اطلاعات متنوعی را به صـــورت یکپارچه در طبقات مختلفی ذخیره نمود تا کاربر نهایی بتولند به نیازهای اطلاعاتی خود در کمترین زمان ممکن پاســـخ دهد. از طرفی دیگر، کتابخلنه ها نیز می‌توانند منابع کاربردی‌تری را از طریق تحلیل علاقه‌مندی های اطلاعاتی کاربران خود تهیه کنند و این نکته در شـرایطی که کتابخانه ها با مشکلات مالی روبرو هستند، می‌تواند نقطه عطفی در نظر گرفته شود.

**کلمات کلیدی:** داده‌کاوی، کتابخانه‌های دیجیتال، مرور نظامند، خدمات دیجیتال، مدیریت دیجیتال، مجموعه‌سازی دیجیتال.

نوع مقاله: مقاله پژوهشی

دریافت: ۱۴۰۰/۰۹/۲۰
پذیرش: ۱۴۰۰/۱۱/۲۳

Contents lists available at ScienceDirect

The Journal of Academic Librarianship

REVIEW

# Literature Review of Data Mining Applications in Academic Libraries

Lorena Siguenza-Guzman [a,b,*], Victor Saquicela [a], Elina Avila-Ordóñez [a,b], Joos Vandewalle [c], Dirk Cattrysse [b]

[a] Department of Computer Science, University of Cuenca, 12 de Abril Av., ECU-010150 Cuenca, Ecuador
[b] Centre for Industrial Management Traffic & Infrastructure, KU Leuven, Celestijnenlaan 300, Box 2422, BE-3001 Leuven, Belgium
[c] Department of Electrical Engineering ESAT/Stadius, KU Leuven, Kasteelpark Arenberg 10, Box 2440, BE-3001 Leuven, Belgium

ARTICLE INFO

*Article history:*
Received 28 January 2015
Accepted 6 June 2015
Available online xxxx

*Keywords:*
Data mining
Bibliomining
Literature review
Academic libraries

ABSTRACT

This article provides a comprehensive literature review and classification method for data mining techniques applied to academic libraries. To achieve this, forty-one practical contributions over the period 1998–2014 were identified and reviewed for their direct relevance. Each article was categorized according to the main data mining functions: clustering, association, classification, and regression; and their application in the four main library aspects: services, quality, collection, and usage behavior. Findings indicate that both collection and usage behavior analyses have received most of the research attention, especially related to collection development and usability of websites and online services respectively. Furthermore, classification and regression models are the two most commonly used data mining functions applied in library settings.

Additionally, results indicate that the top 6 journals of articles published on the application of data mining techniques in academic libraries are: College and Research Libraries, Journal of Academic Librarianship, Information Processing and Management, Library Hi Tech, International Journal of Knowledge, Culture and Change Management, and The Electronic Library. Scopus is the multidisciplinary database that provides the best coverage of journal articles identified. To our knowledge, this study represents the first systematic, identifiable and comprehensive academic literature review of data mining techniques applied to academic libraries.

# Lorena Siguenza-Guzman et al.
# Literature Review of Data Mining Applications in Academic Libraries

**Table 5**
Distribution of articles by year of publication and country of implementation

| | Australia | China | Czech Republic | Germany | Greece | Netherlands | Taiwan | Thailand | UK | USA | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1998 | | | | | | | | | | 1 | 1 |
| 1999 | | | | | | | | | | | 0 |
| 2000 | | | | | | | | | | | 0 |
| 2001 | | | | | | | | | | | 0 |
| 2002 | | | | | | | | | | 2 | 2 |
| 2003 | | | | | 1 | | 3 | | | 2 | 6 |
| 2004 | | | | | | | 1 | | | 1 | 2 |
| 2005 | | | | | | | | | | | 0 |
| 2006 | | | | 2 | | 1 | | | 1 | | 4 |
| 2007 | 1 | | | | | | | | | 1 | 2 |
| 2008 | | | | | | | | | | | 0 |
| 2009 | | | | | | | | | | 2 | 2 |
| 2010 | | | | | | | | | | 3 | 3 |
| 2011 | | | | | | | | 1 | | 3 | 4 |
| 2012 | | | | | 2 | | 2 | | | 2 | 6 |
| 2013 | | 1 | | | | | | | | 1 | 2 |
| 2014 | 1 | 1 | 1 | | | | | | | 4 | 7 |
| Total | 2 | 2 | 1 | 2 | 3 | 1 | 6 | 1 | 1 | 22 | 41 |

**Table 4**
Distribution of articles by data mining techniques and library holistic quadrants[a]

| Data mining techniques | Service analysis | Quality analysis | Collection analysis | Usage analysis | Frequency | Number of articles | Percentage (%) |
|---|---|---|---|---|---|---|---|
| Logistic regression | 4 | | 2 | 2 | 8 | 6 | 15 |
| Association rules | 1 | | 3 | 4 | 8 | 5 | 11 |
| Decision/Classification tree | | | 5 | 1 | 6 | 5 | 11 |
| Logical analysis of data | 3 | | 2 | 2 | 7 | 5 | 11 |
| Linear regression | 4 | 1 | 2 | 4 | 11 | 5 | 11 |
| Log analysis | | | 1 | 4 | 5 | 4 | 9 |
| K-means algorithm | 1 | | | 3 | 4 | 3 | 7 |
| Pattern based clustering | | | 1 | 3 | 4 | 3 | 7 |
| Statistical analysis | | | 1 | 2 | 3 | 3 | 7 |
| Hierarchical cluster analysis | | | | 2 | 2 | 2 | 4 |
| Neutral network | | 2 | 1 | 1 | 4 | 2 | 4 |
| Bibliometric analysis | | | 1 | | 1 | 1 | 2 |
| Memory-based reasoning | | | 1 | | 1 | 1 | 2 |
| Regression analysis | 1 | 1 | | | 2 | 1 | 2 |
| *Count: 14* | *14* | *4* | *20* | *28* | | *Total: 41* | *100.00* |

[a] Remark: each article may have used more than one data mining technique and may have been implemented in more than one library holistic quadrant.

Please cite this article as: Siguenza-Guzman, L., et al., Literature Review of Data Mining Applications in Academic Libraries, *The Journal of Academic Librarianship* (2015), http://dx.doi.org/10.1016/j.acalib.2015.06.007

# Lorena Siguenza-Guzman et al.
# Literature Review of Data Mining Applications in Academic Libraries



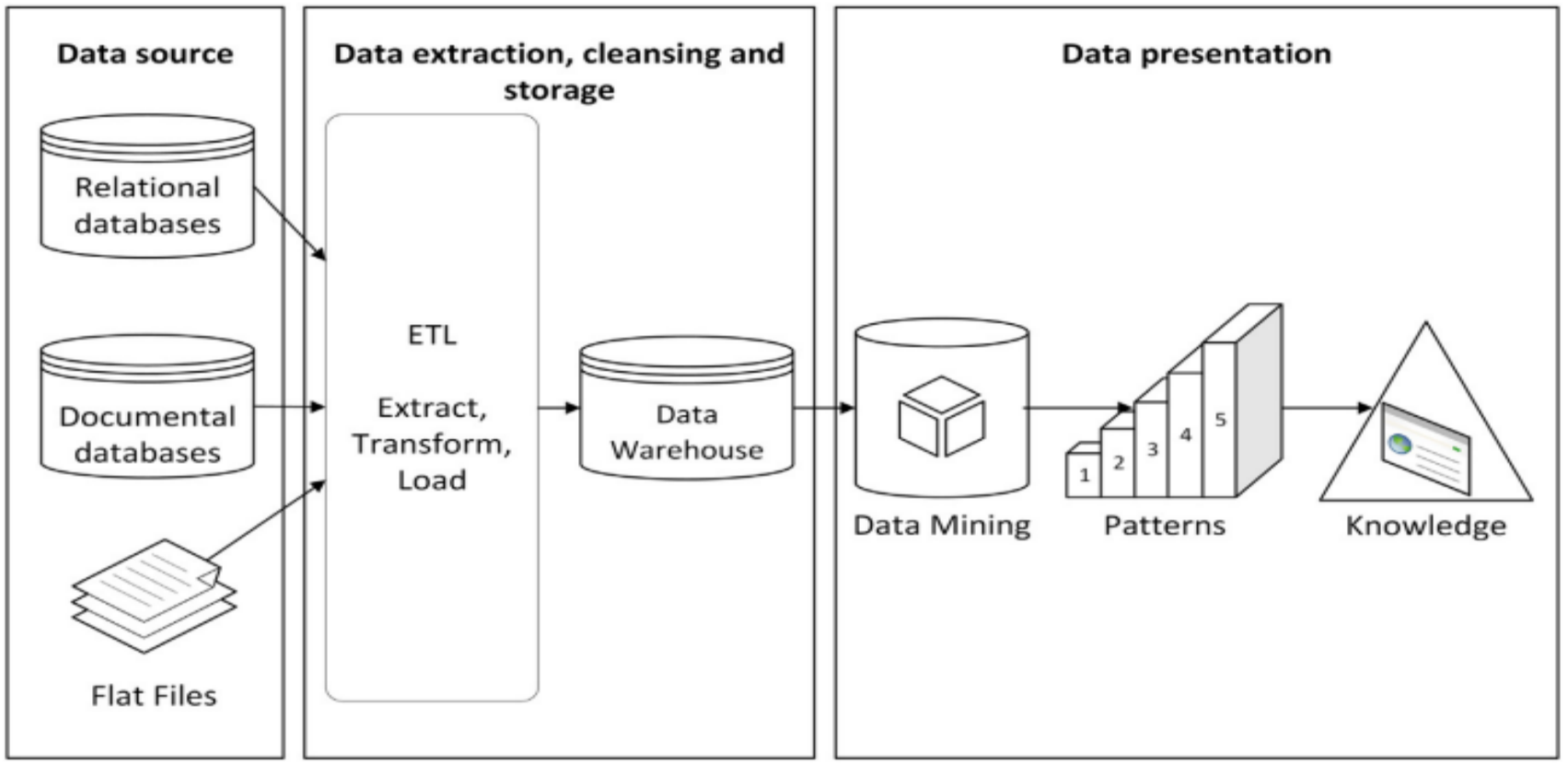



**Fig. 1.** Data mining process, based on Han et al. (2011).

Lorena Siguenza-Guzman, Victor Saquicela, Elina Avila-Ordóñez, Joos Vandewalle, Dirk Cattrysse, Literature Review of Data Mining Applications in Academic Libraries, The Journal of Academic Librarianship,Volume 41, Issue 4,2015,Pages 499-510.

# Lorena Siguenza-Guzman et al.
# Literature Review of Data Mining Applications in Academic Libraries

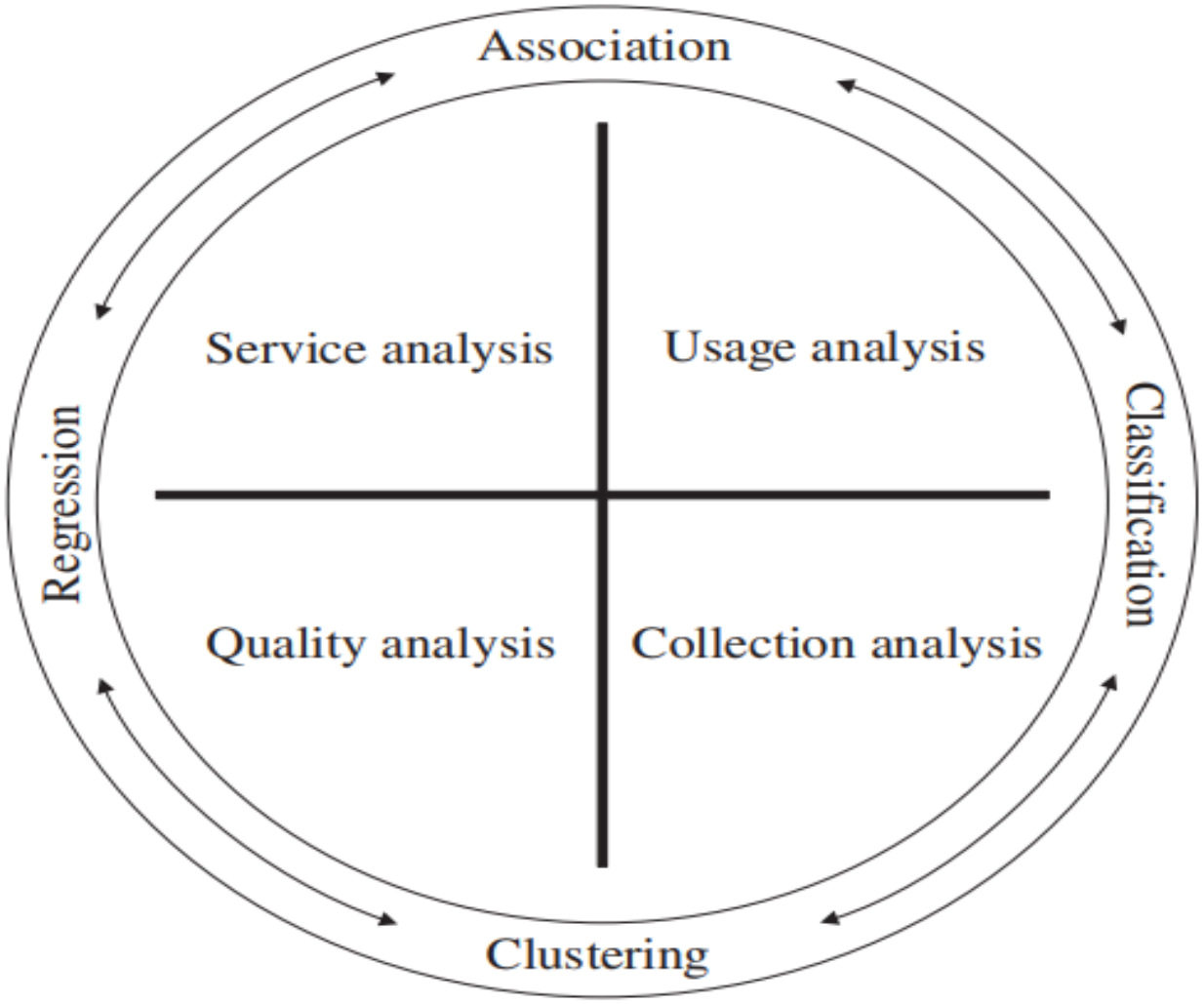



**Fig. 2.** Classification framework for data mining techniques based on the Ngai et al. (2009)

- 4 Data Mining Functions:
  - Clustering
  - Association
  - Classification
  - Regression

- 4 Main Library Aspects:
- Services
- Quality
- Collection
- Usage behavior

- Main Results:
- Most of the research attention:
  - Collection and usage behavior analyses
    - collection development
    - Usability of websites and online services

# Lorena Siguenza-Guzman et al.
# Literature Review of Data Mining Applications in Academic Libraries

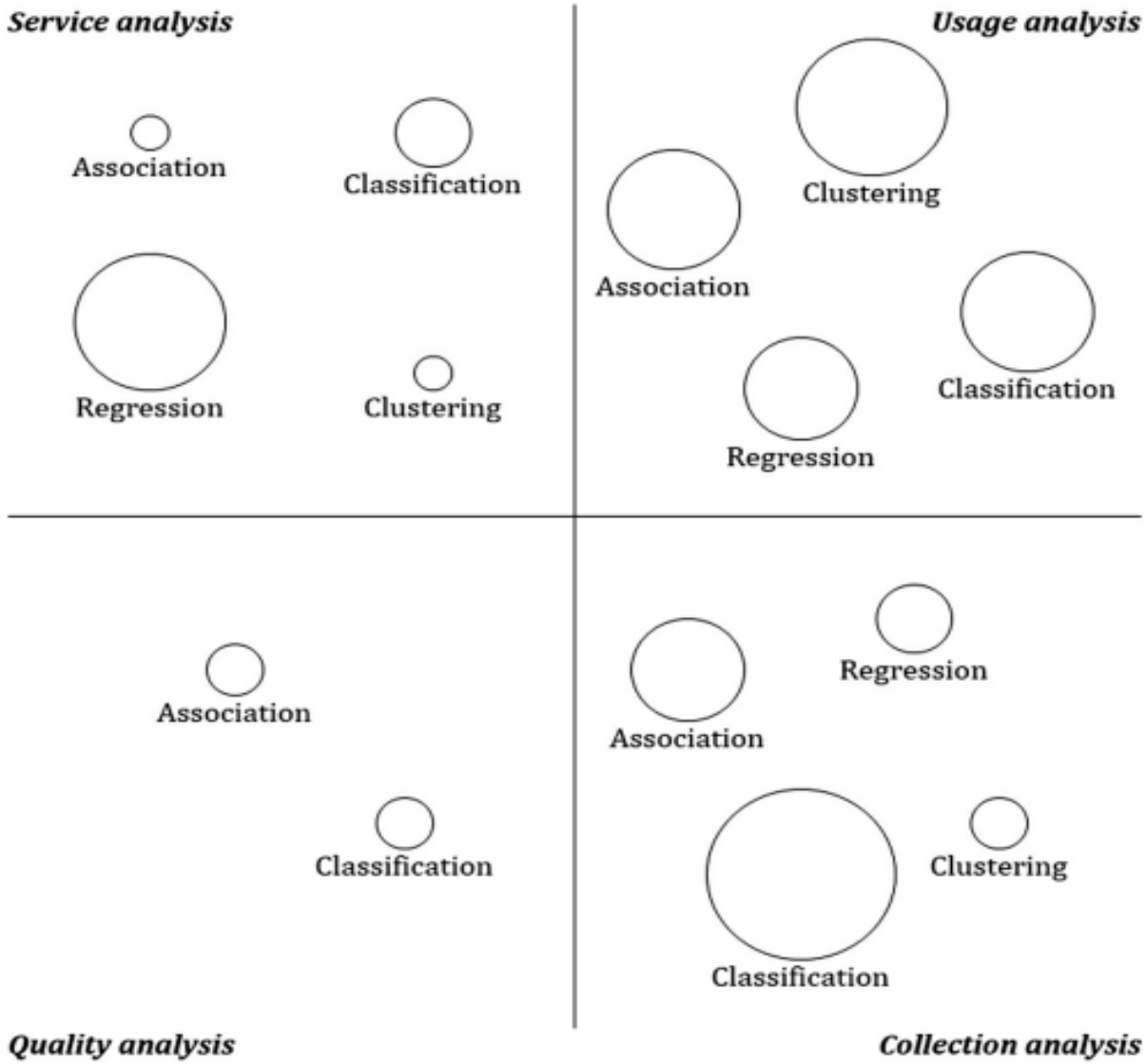



**Fig. 4.** Classification of data mining applications based on the holistic evaluation matrix.

- **Results**
  - the two most commonly used data mining functions applied in library settings
    - Classification Model
    - Regression Models