Обнаружение знаний в базах данных (Knowledge Discovery in Databases)

> Синонимы: Извлечение знаний из баз данных, KDD
> Разделы: Бизнес-задачи
> Loginom: Руководство пользователя

Согласно определению авторов концепции, KDD представляет собой нетривиальный процесс обнаружения корректных, новых, потенциально полезных и интерпретируемых шаблонов в больших массивах данных.

Здесь под данными понимается множество фактов предметной области, представленных в виде записей базы данных, а шаблон — это выражение на некотором языке, описывающее подмножество данных или применяемую к нему модель. Таким образом, поиск шаблонов подразумевает также подгонку моделей к данным, обнаружение в них зависимостей, закономерностей и структур.

«Процесс» означает, что KDD представляет собой многоэтапную, итеративную процедуру, включающую подготовку данных, построение моделей, оценку и уточнение результатов.

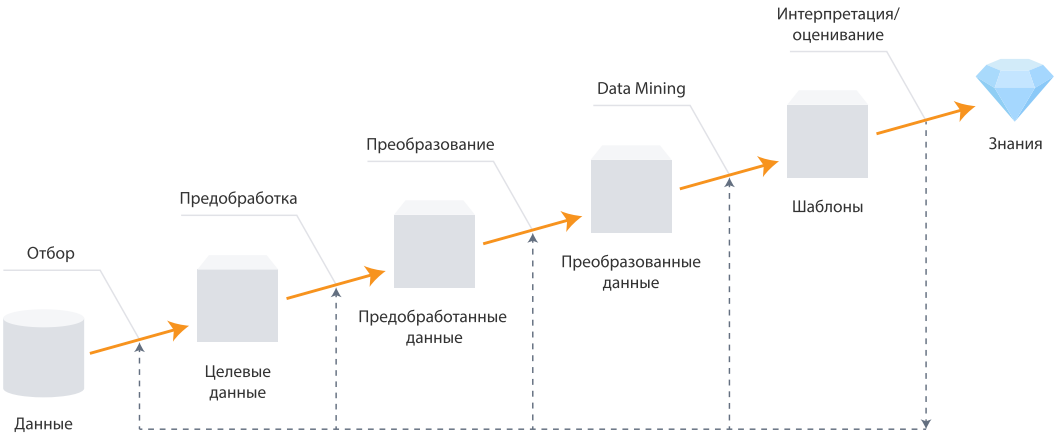
«Нетривиальный» означает, что результаты KDD не должны быть очевидными, а процесс их получения не должен ограничиваться простыми вычислениями, например, средних значений.

Шаблоны, обнаруженные на имеющихся данных, должны быть корректными для любых новых данных предметной области с определенной степенью достоверности. Кроме этого, шаблоны должны были ранее неизвестными и потенциально полезными, то есть позволяющими получить некоторую выгоду при решении определенной задачи. И наконец, шаблоны должны быть понятными и интерпретируемыми, если не сразу, то после некоторой постобработки.

KDD сформировалось и эволюционировало как междисциплинарное направление на стыке таких дисциплин, как:

- машинное обучение;
- распознавание образов;
- базы данных;
- математическая статистика;
- искусственный интеллект;
- экспертные системы;
- визуализация;
- высокопроизводительные вычисления.

Процесс KDD является интерактивным и итеративным, содержащим множество шагов, на каждом из которых пользователь может принимать определенные решения.



В общем случае процесс KDD содержит следующие этапы:

1. **Отбор** — происходит понимание и осмысление предметной области и привлечение априорных знаний о ней, формулирование целей и задач процесса KDD, а также формирование целевого набора данных, на котором будет производиться поиск шаблонов.
2. **Предобработка** — включает очистку данных, отбор данных для построения моделей, выбор методов обработки пропусков и дубликатов, обработку временных рядов и т.д.
3. **Преобразование** (трансформация) — включает сокращение размерности данных, а также определение формы их представления, наиболее оптимальной с точки зрения решаемой задачи.
4. **Data Mining** — к отобранным и подготовленным данным применяются аналитическое методы и модели, решающие задачи классификации и регрессии, поиска ассоциативных правил и кластеризации, а также прогнозирования с целью обнаружения шаблонов.
5. **Интерпретация** — обнаруженные шаблоны представляются в виде решающих правил и их деревьев, структур кластеров, регрессии и т.д. На их основе формируются соответствующие знания. Обнаруженные знания могут быть использованы непосредственно, объединены со знаниями, полученными из других систем и предметных областей, применены для документирования и формирования отчетов. Также на данном этапе производится проверка наличия потенциальных конфликтов с ранее полученными знаниями и их разрешение.

KDD не предписывает, какие методы и алгоритмы обработки следует использовать при решении конкретной задачи, а определяет последовательность действий, которую необходимо выполнить для того, чтобы из исходных данных получить знания. Этот подход универсальный и не зависит от предметной области.

Для поддержки проектов в области KDD в середине 1990-х годов был разработан открытый стандарт моделирования процессов, описывающий общие подходы, используемые при поиске знаний в базах данных, который получил название CRISP-DM.

Основоположниками концепции KDD считаются Пятецкий-Шапиро и Усама Файад (Usama Fayyad).