Добыча данных (Data Mining)

> Синонимы: Разработка данных, Интеллектуальный анализ данных, DM
> Разделы: Бизнес-задачи

Data Mining — это методология и процесс обнаружения в больших массивах данных, накапливающихся в информационных системах компаний, ранее неизвестных, нетривиальных, практически полезных и доступных для интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Data Mining является одним из этапов более масштабной методологии Knowledge Discovery in Databases.

Знания, обнаруженные в процессе Data Mining, должны быть нетривиальными и ранее неизвестными. Нетривиальность предполагает, что такие знания не могут быть обнаружены путем простого визуального анализа. Они должны описывать связи между свойствами бизнес-объектов, предсказывать значения одних признаков на основе других и т.д. Найденные знания должны быть применимы и к новым объектам.

Практическая полезность знаний обусловлена возможностью их использования в процессе поддержки принятия управленческих решений и совершенствовании деятельности компании.

Знания должны быть представлены в виде, понятном для пользователей, которые не имеют специальной математической подготовки. Например, проще всего воспринимаются человеком логические конструкции «если, то». Более того, такие правила могут быть использованы в различных СУБД в качестве SQL-запросов. В случае, когда извлеченные знания непрозрачны для пользователя, должны существовать методы постобработки, позволяющие привести их к интерпретируемому виду.

Data Mining — это не один, а совокупность большого числа различных методов обнаружения знаний. Все задачи, решаемые методами Data Mining, можно условно разбить на шесть видов:

- Классификация;
- Регрессия;
- Кластеризация;
- Ассоциация;
- Последовательные шаблоны;
- Анализ отклонений.

Data Mining носит мультидисциплинарный характер, поскольку включает в себя элементы численных методов, математической статистики и теории вероятностей, теории информации и математической логики, искусственного интеллекта и машинного обучения.

Задачи бизнес-анализа формулируются по-разному, но решение большинства из них сводится к той или иной задаче Data Mining или к их комбинации. Например, оценка рисков — это решение задачи регрессии или классификации, сегментация рынка — кластеризация, стимулирование спроса — ассоциативные правила. Фактически задачи Data Mining являются элементами, из которых можно «собрать» решение большинства реальных бизнес-задач.

Для решения вышеописанных задач используются различные методы и алгоритмы Data Mining. Ввиду того, что Data Mining развивалась и развивается на стыке таких дисциплин, как математическая статистика, теория информации, машинное обучение и базы данных, вполне закономерно, что большинство алгоритмов и методов Data Mining были разработаны на основе различных методов из этих дисциплин. Например, алгоритм кластеризации k-means был заимствован из статистики.

В Data Mining большую популярность получили следующие методы: нейронные сети, деревья решений, алгоритмы кластеризации, в том числе и масштабируемые, алгоритмы обнаружения ассоциативных связей между событиями и т.д.

Основателем и одним из идеологов Data Mining считается Пятецкий-Шапиро. Впервые термин был введен в 1989 году на одном из семинаров, посвященных технологиям поиска знаний в базах данных, проводимых в рамках Международной конференции по искусственному интеллекту (International Joint Conference on Artificial Intelligence) IJCAI-89.

В разделе Data Mining описаны все обработчики в Loginom, относящиеся к данному направлению обработки информации. В статье «Алгоритмы кластеризации на службе Data Mining» описываются алгоритмы кластеризации с точки зрения их применения в Data Mining.