Лекция 1. МЕТОДОЛОГИИ И ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ

Понятие данных. Данные – это воспринимаемые человеком факты, события, сообщения, измеряемые характеристики, регистрируемые сигналы.

Анализ данных представляет собой процесс изучения статистических данных (поиска статистических закономерностей, закономерностей в среднем) с помощью математических методов, не предполагающих вероятностной модели изучаемого явления. Противостоит вероятностно-статистическому подходу к обработке данных, опирающемуся на их вероятностную интерпретацию (как случайной выборки из генеральной совокупности) и использование вероятностных моделей для построения и выбора наилучших методов обработки. Получаемые с помощью вероятностно-статистического подхода выводы опираются на строго доказанные математические положения. В частности, этот подход обеспечивает корректный перенос результатов с выборки на генеральную совокупность. В методах анализа данных подобные возможности не заложены. Эти методы не удовлетворяют строгим математическим требованиям. Выбор наилучшего метода здесь почти всегда опирается на неформализуемые эвристические соображения. Поэтому проблема обоснования получаемых выводов требует особого внимания. Особенно острой становится необходимость выделения «точек соприкосновения» содержания задачи и математического формализма, реализации в процессе человеко-машинного диалога.

Анализ данных — область математики и информатики, занимающаяся построением и исследованием наиболее общих математических методов и вычислительных алгоритмов извлечения знаний из экспериментальных (в широком смысле) данных;

Интеллектуальный анализ данных (Data Mining) — это особый метод анализа данных, который фокусируется на моделировании и открытии данных, а не на их описании. Бизнес-аналитика охватывает анализ данных, который полагается на агрегацию. В статистическом смысле некоторые разделяют анализ данных на описательную статистику, исследовательский анализ данных и проверку статистических гипотез. Исследовательский анализ данных занимается открытием новых характеристик данных, а проверка статистических гипотез — на подтверждении или опровержении

существующих гипотез. Прогнозный анализ фокусируется на применении статистических или структурных моделей для предсказания или классификации, а анализ текста применяет статистические, лингвистические и структурные методы для извлечения и классификации информации из текстовых источников, принадлежащих к неструктурированным данным. Все это разновидности анализа данных.

В общих чертах внедрение данных в систему и эксплуатацию полученной модели можно представить в виде блок-схемы, изображенной на рис.1.1.

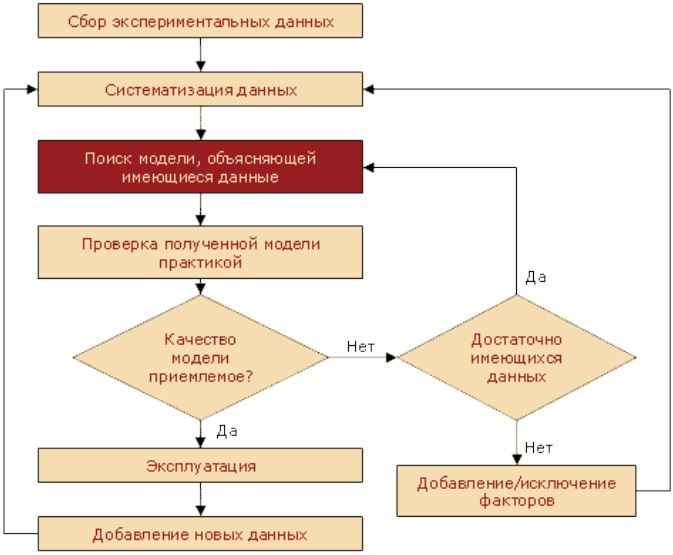

Рис.1.1 Общая схема работы с данными

Этапы анализа данных. Выделяют четыре основных этапа анализа данных:

1. Описание совокупности данных. Целью данного этапа является поиск "выбросов" (неправильно введенных ответов респондента) и логических нарушений в ходе интервью (например, не сделанный переход). Выделяются следующие подэтапы:

1) Чистка массива

- выявление ошибок и пропусков, допущенных в ходе сбора и ввода информации.

- коррекция выборки. Наиболее распространенным методом коррекции выборки является перевзвешивание. При его использовании, ответы более представленной категории респондентов учитываются с определенным коэффициентом (например, 0,9).

2) Описание. Включает в себя описание распределения данных по существенным с точки зрения целей и задачи признакам. Здесь используются следующие методы: статистической группировки – строятся одномерные распределения; меры центральной тенденции; меры рассеяния, вариации.

2. Уплотнение исходной информации. Целью данного этапа является сокращение числа признаков, необходимых для анализа. Выделяются следующие подэтапы:

- укрупнение шкал (например, группировка возраста)
- расчет индексов и агрегированных показателей.

3. Углубление интерпретации и переход к объяснению. Целью данного этапа является поиск статистических закономерностей в распределении данных. Здесь же проверяются основные гипотезы, строятся выводы. На данном этапе основной применяемый метод – корреляционный анализ.

4. Прогноз развития явлений. Этот этап имеет место лишь в аналитических исследованиях. Происходит построение содержательных представлений об основе процесса.

Методы анализа данных. Все методы подразделяются на две большие группы по принципу работы с исходными обучающими данными. В этой классификации верхний уровень определяется на основании того, сохраняются ли данные после Data Mining либо они дистиллируются для последующего использования.

1. Непосредственное использование данных, или сохранение данных. В этом случае исходные данные хранятся в явном детализированном виде и непосредственно используются на стадиях прогностического моделирования и/или анализа исключений. Проблема этой группы методов - при их использовании могут возникнуть сложности анализа сверхбольших баз данных. Методы этой группы: кластерный анализ, метод ближайшего соседа, метод k-ближайшего соседа, рассуждение по аналогии.

2. Выявление и использование формализованных закономерностей, или дистилляция шаблонов. При технологии дистилляции шаблонов один образец (шаблон) информации извлекается из исходных данных и преобразуется в некие формальные конструкции, вид которых зависит от используемого метода. Этот

процесс выполняется на стадии свободного поиска, у первой же группы методов данная стадия в принципе отсутствует. На стадиях прогностического моделирования и анализа исключений используются результаты стадии свободного поиска, они значительно компактнее самих баз данных. Конструкции этих моделей могут быть трактуемыми аналитиком либо нетрактуемыми ("черными ящиками"). Методы этой группы: логические методы; методы визуализации; методы кросс-табуляции; методы, основанные на уравнениях.

Другая классификация разделяет все многообразие методов на две группы: статистические и кибернетические методы. Эта схема разделения основана на различных подходах к обучению математических моделей

Статистические методы. Все методы представляют собой четыре взаимосвязанных раздела:

- предварительный анализ природы статистических данных (проверка гипотез стационарности, нормальности, независимости, однородности, оценка вида функции распределения, ее параметров и т.п.);
- выявление связей и закономерностей (линейный и нелинейный регрессионный анализ, корреляционный анализ и др.);
- многомерный статистический анализ (линейный и нелинейный дискриминантный анализ, кластерный анализ, компонентный анализ, факторный анализ и др.);
- динамические модели и прогноз на основе временных рядов.

Арсенал статистических методов классифицирован на четыре группы методов:

1. Дескриптивный анализ и описание исходных данных.
2. Анализ связей (корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ).
3. Многомерный статистический анализ (компонентный анализ, дискриминантный анализ, многомерный регрессионный анализ, канонические корреляции и др.).
4. Анализ временных рядов (динамические модели и прогнозирование).

Кибернетические методы. Второе направление - это множество подходов, объединенных идеей компьютерной математики и использования теории искусственного интеллекта.

К этой группе относятся такие методы:

- искусственные нейронные сети (распознавание, кластеризация, прогноз);

- эволюционное программирование (в т.ч. алгоритмы метода группового учета аргументов);
- генетические алгоритмы (оптимизация);
- ассоциативная память (поиск аналогов, прототипов);
- нечеткая логика;
- деревья решений;
- системы обработки экспертных знаний.

Задачи анализа данных. В настоящее время выделяют следующие основные классы задач анализа данных:

1. Прогнозирование (Forecasting). Нахождение будущих состояний объекта на основании предыдущих состояний (исторических данных).

Примеры:

- прогнозирование ситуаций на валютных рынках
- прогнозирование цен на рынке недвижимости
- прогнозирование демографических процессов
- прогнозирование климатических процессов

2. Классификация (Classification). Нахождение правила, позволяющее отнести объект к тому или иному классу (выбрать класс из числа известных заранее классов) на основе информации о том, к какому классу относятся другие объекты.

Примеры:

- Задачи распознавание образов (распознавание рукописного текста, фотографии, идентификация личности по фото, голосу, видео…)
- Задачи атрибуции (определение авторства / периода создания / страны происхождения … произведений искусства, археологических находок)
- Задачи диагностики (в медицине и технике).

3. Кластеризация (Clusterization). Нахождение правила для автоматического разделения имеющихся объектов на классы на основании сходства тех или иных характеристик (факторов) этих объектов. При этом ни сами классы, ни их количество заранее неизвестны.

Примеры:

- Сегментация рынка (разделение всех потенциальных потребителей на кластеры для последующего целевого воздействия, например, создания целевой рекламы)

- Задачи разбиения множества индивидов на группы кластеры (в социологии, психологии, биологии и пр…)

4. Ассоциация (Associations). Поиск устойчивых закономерностей между случайными событиями, наступающими одновременно.

Пример:

- Анализ покупательской корзины – поиск «устойчивых связей в корзине покупателя» (осуществляется с целью учёта их при планировании расположения отделов в супермаркете).

5. Последовательность (Последовательная ассоциация, Нахождение последовательных шаблонов) (Sequence, Sequential association, Sequential pattern). Поиск устойчивых закономерностей между случайными событиями, связанными во времени, т.е. правил вида: после события X через время t происходит событие Y.

Пример:

- После покупки квартиры жильцы в 60% случаев в течение двух недель приобретают холодильник, а в течение двух месяцев в 50% случаев приобретается телевизор. Решение данной задачи широко применяется в маркетинге и менеджменте, например, при управлении циклом работы с клиентом (Customer Lifecycle Management).
- Визуализация данных (Data Visualization). Графическое изображение данных (2D и 3D диаграммы, гистограммы, графики, облака точек…)

6. Анализ отклонений (Deviation Detection). Обнаружение и анализ данных, наиболее отличающихся от общего множества данных.

Примеры:

- выявление нетипичной сетевой активности позволяет обнаружить вредоносные программы
- выявление мошенничества с кредитными карточками.

Используемые технологии

С точки зрения типа данных, которые используются для анализа, сегодня можно выделить:

- Data Mining,
- Text Mining
- Visual Mining
- OLAP,
- анализ процессов (Process Mining),
- анализ Web-ресурсов (Web mining)

- и анализ в режиме реального времени (Real-Time Data Mining).

Data Mining. Data Mining – это процесс обнаружения в "сырых" данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Data Mining является одним из шагов Knowledge Discovery in Databases.

Информация, найденная в процессе применения методов Data Mining, должна быть нетривиальной и ранее неизвестной, практически полезной и доступной интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Основу методов Data Mining составляют всевозможные методы классификации, моделирования и прогнозирования, основанные на применении деревьев решений, искусственных нейронных сетей, генетических алгоритмов, эволюционного программирования, ассоциативной памяти, нечёткой логики. К методам Data Mining нередко относят статистические методы (дескриптивный анализ, корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ, компонентный анализ, дискриминантный анализ, анализ временных рядов, анализ выживаемости, анализ связей). Такие методы, однако, предполагают некоторые априорные представления об анализируемых данных, что несколько расходится с целями Data Mining (обнаружение ранее неизвестных нетривиальных и практически полезных знаний).

Одно из важнейших назначений методов Data Mining состоит в наглядном представлении результатов вычислений (визуализация), что позволяет использовать инструментарий Data Mining людьми, не имеющими специальной математической подготовки.

Методы Data Mining могут быть применены как для работы с большими данными, так и для обработки сравнительно малых объемов данных (полученных, например, по результатам отдельных экспериментов, либо при анализе данных о деятельности компании). В качестве критерия достаточного количества данных рассматривается как область исследования, так и применяемый алгоритм анализа

Text Mining. Интеллектуальный анализ текстов (ИАТ, англ. text mining) — направление в искусственном интеллекте, целью которого является получение информации из коллекций текстовых документов, основываясь на применении эффективных в практическом плане методов машинного обучения и обработки естественного языка. Название «интеллектуальный анализ текстов» перекликается с понятием «интеллектуальный анализ данных» (ИАД, англ. data

mining), что выражает схожесть их целей, подходов к переработке информации и сфер применения; разница проявляется лишь в конечных методах, а также в том, что ИАД имеет дело с хранилищами и базами данных, а не электронными библиотеками и корпусами текстов.

Ключевыми группами задач ИАТ являются: категоризация текстов, извлечение информации и информационный поиск, обработка изменений в коллекциях текстов, а также разработка средств представления информации для пользователя.

Категоризация документов заключается в отнесении документов из коллекции к одной или нескольким группам (классам, кластерам) схожих между собой текстов (например, по теме или стилю). Категоризация может происходить при участии человека, так и без него. В первом случае, называемом классификацией документов, система ИАТ должна отнести тексты к уже определённым (удобным для него) классам. В терминах машинного обучения для этого необходимо произвести обучение с учителем, для чего пользователь должен предоставить системе ИАТ как множество классов, так и образцы документов, принадлежащих этим классам.

Второй случай категоризации называется кластеризацией документов. При этом система ИАТ должна сама определить множество кластеров, по которым могут быть распределены тексты, — в машинном обучении соответствующая задача называется обучением без учителя. В этом случае пользователь должен сообщить системе ИАТ количество кластеров, на которое ему хотелось бы разбить обрабатываемую коллекцию (подразумевается, что в алгоритм программы уже заложена процедура выбора признаков).

Visual Mining. VDM - процесс извлечения скрытых, не выраженных явным образом полезных знаний из больших наборов данных (data set); помогает специалистам обнаруживать новые тенденции, выявлять скрытые связи и закономерности в массивах данных, представленных в самых разных форматах. Состоит из четырёх направлений: визуализации данных (data visualization), визуализации результатов извлечения знаний из данных (mining result visualization), визуализации процесса извлечения знаний из данных (mining process visualization) и интерактивного визуального анализа данных (interactive visual mining) с помощью методов визуализации

За счёт того, что пользователь напрямую работает с данными, представленными в виде визуальных образов, которые он может рассматривать с разных сторон и под любыми углами зрения, в

прямом смысле этого слова, он может получить дополнительную информацию, которая поможет ему более чётко сформулировать цели исследования.

Таким образом, визуальный анализ данных можно представить как процесс генерации гипотез. При этом сгенерированные гипотезы можно проверить или автоматическими средствами (методами статистического анализа или методами Data Mining), или средствами визуального анализа.

Кроме того, прямое вовлечение пользователя в визуальный анализ имеет два основных преимущества перед автоматическими методами:

- визуальный анализ данных позволяет легко работать с неоднородными и зашумлёнными данными, в то время как не все автоматические методы могут работать с такими данными и давать удовлетворительные результаты;
- визуальный анализ данных интуитивно понятен и не требует сложных математических или статистических алгоритмов.

Визуальный анализ данных обычно выполняется в три этапа:

- беглый анализ - позволяет идентифицировать интересные шаблоны и сфокусироваться на одном или нескольких из них;
- увеличение и фильтрация - идентифицированные на предыдущем этапе шаблоны отфильтровываются и рассматриваются в большем масштабе;
- детализация по необходимости - если пользователю нужно получить дополнительную информацию, он может визуализировать более детальные данные

OLAP. OLAP (англ. online analytical processing, интерактивная аналитическая обработка) — технология обработки данных, заключающаяся в подготовке суммарной (агрегированной) информации на основе больших массивов данных, структурированных по многомерному принципу. Реализации технологии OLAP являются компонентами программных решений класса Business Intelligence.

OLAP-куб содержит базовые данные и информацию об измерениях (агрегаты). Куб потенциально содержит всю информацию, которая может потребоваться для ответов на любые запросы. При огромном количестве агрегатов зачастую полный расчёт происходит только для некоторых измерений, для остальных же производится «по требованию».

Существуют три типа OLAP:

- многомерная OLAP (Multidimensional OLAP — MOLAP);

- реляционная OLAP (Relational OLAP — ROLAP);
- гибридная OLAP (Hybrid OLAP — HOLAP).

MOLAP — классическая форма OLAP, так что её часто называют просто OLAP. Она использует суммирующую БД, специальный вариант процессора пространственных БД и создаёт требуемую пространственную схему данных с сохранением как базовых данных, так и агрегатов.

ROLAP работает напрямую с реляционным хранилищем, факты и таблицы с измерениями хранятся в реляционных таблицах, и для хранения агрегатов создаются дополнительные реляционные таблицы.

HOLAP использует реляционные таблицы для хранения базовых данных и многомерные таблицы для агрегатов.

Process Mining. Интеллектуальный анализ процессов (Process Mining) фокусируется на обнаружении, анализе и оптимизации бизнес-процессов на основе данных из журналов событий (англ. event logs), представляя недостающее звено между классическим анализом бизнес-процессов с использованием их моделей и интеллектуальным анализом данных

Как правило, методы процессной аналитики используются в тех случаях, когда формальное описание или модель системы отсутствуют или имеют низкое соответствие реальному поведению системы. Для построения модели могут использоваться журналы функционирования информационных систем (англ. workflow management system, ERP), организаций и предприятий.

Методы процессной аналитики подразделяют на три главных раздела:

- Обнаружение (Discovery): Модель процесса строится на основании журнала событий, содержащего запись функционирования информационной системы за некоторый промежуток времени.
- Проверка соответствия (Conformance checking): Существующая модель сравнивается с журналом событий, анализируются выявленные несоответствия в поведении реальной системы и моделируемом поведении.
- Усовершенствование (Enhancement): Существующая модель усовершенствуется путём расширения моделируемого поведения или повышения эффективности моделирования с использованием информации, полученной в ходе проверки соответствия модели и журнала событий.

Web Mining. Web Mining можно перевести как "добыча данных в Web". Web Intelligence или Web Интеллект готов "открыть новую

главу" в стремительном развитии электронного бизнеса. Способность определять интересы и предпочтения каждого посетителя, наблюдая за его поведением, является серьезным и критичным преимуществом конкурентной борьбы на рынке электронной коммерции. Системы Web Mining могут ответить на многие вопросы, например, кто из посетителей является потенциальным клиентом Web-магазина, какая группа клиентов Web-магазина приносит наибольший доход, каковы интересы определенного посетителя или группы посетителей.

Технология Web Mining охватывает методы, которые способны на основе данных сайта обнаружить новые, ранее неизвестные знания и которые в дальнейшем можно будет использовать на практике. Другими словами, технология Web Mining применяет технологию Data Mining для анализа неструктурированной, неоднородной, распределенной и значительной по объему информации, содержащейся на Web-узлах.

Согласно таксономии Web Mining, здесь можно выделить два основных направления: Web Content Mining и Web Usage Mining.

Web Content Mining подразумевает автоматический поиск и извлечение качественной информации из разнообразных источников Интернета, перегруженных "информационным шумом". Здесь также идет речь о различных средствах кластеризации и аннотировании документов. В этом направлении, в свою очередь, выделяют два подхода: подход, основанный на агентах, и подход, основанный на базах данных.

Подход, основанный на агентах (Agent Based Approach), включает такие системы:

- интеллектуальные поисковые агенты (Intelligent Search Agents);
- фильтрация информации / классификация;
- персонифицированные агенты сети.

Подход, основанный на базах данных (Database Approach), включает системы:

- многоуровневые базы данных;
- системы web-запросов (Web Query Systems);

Второе направление Web Usage Mining подразумевает обнаружение закономерностей в действиях пользователя Web-узла или их группы.

Анализируется следующая информация:

- какие страницы просматривал пользователь;
- какова последовательность просмотра страниц.

Анализируется также, какие группы пользователей можно выделить среди общего их числа на основе истории просмотра Web-узла.

Web Usage Mining включает следующие составляющие:

- предварительная обработка;
- операционная идентификация;
- инструменты обнаружения шаблонов;
- инструменты анализа шаблонов.

При использовании Web Mining перед разработчиками возникает два типа задач. Первая касается сбора данных, вторая - использования методов персонификации. В результате сбора некоторого объема персонифицированных ретроспективных данных о конкретном клиенте, система накапливает определенные знания о нем и может рекомендовать ему, например, определенные наборы товаров или услуг. На основе информации о всех посетителях сайта Web-система может выявить определенные группы посетителей и также рекомендовать им товары или же предлагать товары в рассылках.

Задачи Web Mining можно подразделить на такие категории:

- Предварительная обработка данных для Web Mining.
- Обнаружение шаблонов и открытие знаний с использованием ассоциативных правил, временных последовательностей, классификации и кластеризации;
- Анализ полученного знания.

Real-Time Data Mining. Методы Data Mining в реальном времени (или Real-Time Analytics), в основном, относятся к задаче предсказания. В отличие от статических методов ониобучаются динамически и основаны на обратной связи от прогноза, полученного с помощью предсказательной модели (постоянном переобучении).

Интеллектуальный анализ данных в реальном времени определяется как набор методов, имеющих все следующие характеристики, не зависящие от объема используемых данных:

1. Инкрементальное обучение (обучение)
2. Декрементальное обучение (забыть)
3. Добавление атрибутов (рост)
4. Удаление атрибутов (сокращение)
5. Распределенная обработка.
6. Параллельная обработка.

Обновление обычного интеллектуального анализа данных до интеллектуального анализа данных в реальном времени осуществляется с помощью метода, называемого «Учебная машина в

реальном времени» или RTLM. Использование RTLM с традиционными методами интеллектуального анализа данных обеспечивает «интеллектуальный анализ данных в реальном времени».

Сферы применения. Сейчас технология Data Mining используется практически во всех сферах деятельности человека, где накоплены ретроспективные данные:

- Применение Data Mining для решения бизнес-задач. Основные направления: банковское дело, финансы, страхование, CRM, производство, телекоммуникации, электронная коммерция, маркетинг, фондовый рынок и другие.
- Применение Data Mining для решения задач государственного уровня. Основные направления: поиск лиц, уклоняющихся от налогов; средства в борьбе с терроризмом.
- Применение Data Mining для научных исследований. Основные направления: медицина, биология, молекулярная генетика и генная инженерия, биоинформатика, астрономия, прикладная химия, исследования, касающиеся наркотической зависимости, и другие.
- Применение Data Mining для решения Web-задач. Основные направления: поисковые машины (search engines), счетчики и другие.

Применение Data Mining для решения бизнес-задач

- Банковское дело. Технология Data Mining используется в банковской сфере для решения ряда типичных задач. Классический пример применения Data Mining в банковском деле - решение задачи определения возможной некредитоспособности клиента банка. Эту задачу также называют анализом кредитоспособности клиента или "Выдавать ли кредит клиенту?". Без применения технологии Data Mining задача решается сотрудниками банковского учреждения на основе их опыта, интуиции и субъективных представлений о том, какой клиент является благонадежным. По похожей схеме работают системы поддержки принятия решений и на основе методов Data Mining. Такие системы на основе исторической (ретроспективной) информации и при помощи методов классификации выявляют клиентов, которые в прошлом не вернули кредит.
- Страхование. Страховой бизнес связан с определенным риском. Здесь задачи, решаемые при помощи Data Mining, сходны с задачами в банковском деле. Информация, полученная в результате сегментации клиентов на группы, используется для определения групп клиентов. В результате страховая компания может с

наибольшей выгодой и наименьшим риском предлагать определенные группы услуг конкретным группам клиентов.

- Телекоммуникации. В сфере телекоммуникаций достижения Data Mining могут использоваться для решения задачи, типичной для любой компании, которая работает с целью привлечения постоянных клиентов, - определения лояльности этих клиентов. Необходимость решения таких задач обусловлена жесткой конкуренцией на рынке телекоммуникаций и постоянной миграцией клиентов от одной компании в другую. Как известно, удержание клиента намного дешевле его возврата. Поэтому возникает необходимость выявления определенных групп клиентов и разработка наборов услуг, наиболее привлекательных именно для них. В этой сфере, так же как и во многих других, важной задачей является выявление фактов мошенничества.
- Электронная коммерция. В сфере электронной коммерции Data Mining применяется для формирования рекомендательных систем и решения задач классификации посетителей Web-сайтов. Такая классификация позволяет компаниям выявлять определенные группы клиентов и проводить маркетинговую политику в соответствии с обнаруженными интересами и потребностями клиентов. Технология Data Mining для электронной коммерции тесно связана с технологией Web Mining [28].
- Промышленное производство. Особенности промышленного производства и технологических процессов создают хорошие предпосылки для возможности использования технологии Data Mining в ходе решения различных производственных задач. Технический процесс по своей природе должен быть контролируемым, а все его отклонения находятся в заранее известных пределах;
- Применение Data Mining в CRM. Одно из наиболее перспективных направлений применения Data Mining - использование данной технологии в аналитическом CRM. CRM (Customer Relationship Management) - управление отношениями с клиентами. При совместном использовании этих технологий добыча знаний совмещается с "добычей денег" из данных о клиентах. Важным аспектом в работе отделов маркетинга и отдела продаж является составление целостного представления о клиентах, информация об их особенностях, характеристиках, структуре клиентской базы. В CRM используется так называемое профилирование клиентов, дающее полное представление всей необходимой информации о клиентах. Профилирование клиентов включает следующие компоненты:

сегментация клиентов, прибыльность клиентов, удержание клиентов, анализ реакции клиентов. Каждый из этих компонентов может исследоваться при помощи Data Mining, а анализ их в совокупности, как компонентов профилирования, в результате может дать те знания, которые из каждой отдельной характеристики получить невозможно.

В результате использования Data Mining решается задача сегментации клиентов на основе их прибыльности. Анализ выделяет те сегменты покупателей, которые приносят наибольшую прибыль. Сегментация также может осуществляться на основе лояльности клиентов. В результате сегментации вся клиентская база будет поделена на определенные сегменты, с общими характеристиками. В соответствии с этими характеристиками компания может индивидуально подбирать маркетинговую политику для каждой группы клиентов.

Также можно использовать технологию Data Mining для прогнозирования реакции определенного сегмента клиентов на определенный вид рекламы или рекламных акций - на основе ретроспективных данных, накопленных в предыдущие периоды.

Таким образом, определяя закономерности поведения клиентов при помощи технологии Data Mining, можно существенно повысить эффективность работы отделов маркетинга, продаж и сбыта. При объединении технологий CRM и Data Mining и грамотном их внедрении в бизнес компания получает значительные преимущества перед конкурентами.

Data Mining для научных исследований

- Биоинформатика. Одна из научных областей применения технологии Data Mining - биоинформатика, направление, целью которого является разработка алгоритмов для анализа и систематизации генетической информации. Полученные алгоритмы используются для определения структур макромолекул, а также их функций, с целью объяснения различных биологических явлений.
- Медицина. Несмотря на консервативность медицины во многих ее аспектах, технология Data Mining в последние годы активно применяется для различных исследований и в этой сфере человеческой деятельности. Традиционно для постановки медицинских диагнозов используются экспертные системы, которые построены на основе символьных правил, сочетающих, например, симптомы пациента и его заболевание. С использованием Data Mining при помощи шаблонов можно разработать базу знаний для экспертной системы.

- Фармацевтика. В области фармацевтики методы Data Mining также имеют достаточно широкое применение. Это задачи исследования эффективности клинического применения определенных препаратов, определение групп препаратов, которые будут эффективны для конкретных групп пациентов. Актуальными здесь также являются задачи продвижения лекарственных препаратов на рынок.
- Молекулярная генетика и генная инженерия. В молекулярной генетике и генной инженерии выделяют отдельное направление Data Mining, которое имеет название анализ данных в микро-массивах (Microarray Data Analysis, MDA). Примеры использования Data Mining - молекулярный диагноз некоторых серьезнейших заболеваний; открытие того, что генетический код действительно может предсказывать вероятность заболевания; открытие некоторых новых лекарств и препаратов. Основные понятия, которыми оперирует Data Mining в областях "Молекулярная генетика и генная инженерия" - маркеры, т.е. генетические коды, которые контролируют различные признаки живого организма.
- Химия. Технология Data Mining активно используется в исследованиях органической и неорганической химии. Одно из возможных применений Data Mining в этой сфере - выявление каких-либо специфических особенностей строения соединений, которые могут включать тысячи элементов.