2 ДАННЫЕ

В широком понимании *данные* представляют собой факты, текст, графики, картинки, звуки, аналоговые или цифровые видео-сегменты.

Данные могут быть получены в результате измерений, экспериментов, арифметических и логических операций.

Данные должны быть представлены в форме, пригодной для хранения, передачи и обработки.

Иными словами, ***данные*** - это необработанный материал, предоставляемый поставщиками *данных* и используемый потребителями для формирования информации на основе *данных*.

Набор данных и их атрибутов

В таблице 2.1 представлена двухмерная *таблица*, представляющая собой набор *данных*.

Таблица 2.1. Двухмерная таблица "объект-атрибут"

	Атрибуты				
Объекты	Код клиента	Возраст	Семейное положение	Доход	Класс
	1	18	Single	125	1
	2	22	Married	100	1
	3	30	Single	70	1
	4	32	Married	120	1
	5	24	Divorced	95	2
	6	25	Married	60	1
	7	32	Divorced	220	1
	8	19	Single	85	2
	9	22	Married	75	1
	10	40	Single	90	2

По горизонтали таблицы располагаются *атрибуты* объекта или его признаки. *По* вертикали таблицы - *объекты*.

Объект описывается как набор атрибутов.

Объект также известен как *запись*, случай, пример, строка таблицы и т.д.

Атрибут - свойство, характеризующее *объект*.

Например: цвет глаз человека, температура воды и т.д.

Атрибут также называют переменной, полем таблицы, измерением, характеристикой.

В результате операционализации понятий [6], т.е. перехода от общих категорий к конкретным величинам, получается набор переменных изучаемого понятия.

Переменная (variable) - свойство или характеристика, общая для всех изучаемых *объектов*, проявление которой может изменяться от *объекта* к *объекту*.

Значение (value) переменной является проявлением признака.

При анализе *данных*, как правило, нет возможности рассмотреть всю интересующую нас совокупность *объектов*. Изучение очень больших объемов *данных* является дорогостоящим процессом, требующим больших временных затрат, а также неизбежно приводит к ошибкам, связанным с человеческим фактором.

Вполне достаточно рассмотреть некоторую часть всей совокупности, то есть *выборку*, и получить интересующую нас информацию на ее основании.

Однако размер *выборки* должен зависеть от разнообразия *объектов*, представленных в генеральной совокупности. В *выборке* должны быть представлены различные комбинации и элементы генеральной совокупности.

Генеральная совокупность (*population*) - вся совокупность изучаемых *объектов*, интересующая исследователя.

Выборка (sample) - часть генеральной совокупности, определенным способом отобранная с целью исследования и получения выводов о свойствах и характеристиках генеральной совокупности.

Параметры - числовые характеристики генеральной совокупности.

Статистики - числовые характеристики *выборки*.

Часто исследования основываются на *гипотезах*. *Гипотезы* проверяются с помощью *данных*.

Гипотеза - предположение относительно параметров совокупности объектов, которое должно быть проверено на ее части.

Гипотеза - частично обоснованная *закономерность* знаний, служащая либо для связи между различными эмпирическими фактами, либо для объяснения факта или группы фактов.

Пример *гипотезы*: между показателями продолжительности жизни и качеством питания есть *связь*. В этом случае целью исследования может быть объяснение изменений конкретной переменной, в данном случае - продолжительности жизни. Допустим, существует *гипотеза*, что **зависимая переменная** (продолжительность жизни) изменяется в **зависимости** от некоторых причин (качество питания, образ жизни, *место* проживания и т.д.), которые и являются **независимыми переменными.**

Однако *переменная* изначально не является зависимой или независимой. Она становится таковой после формулировки конкретной *гипотезы*. Зависимая *переменная* в одной *гипотезе* может быть независимой в другой.

Измерения

Измерение - процесс присвоения чисел характеристикам изучаемых *объектов* согласно определенному правилу.

В процессе подготовки *данных* измеряется не сам *объект*, а его характеристики.

Шкала - правило, в соответствии с которым объектам присваиваются числа.

Многие инструменты *Data* Mining при импорте *данных* из других источников предлагают выбрать тип *шкалы* для каждой переменной и/или выбрать тип *данных* для входных и выходных переменных (символьные, числовые, дискретные и непрерывные). Пользователю такого инструмента необходимо владеть этими понятиями.

Переменные могут являться **числовыми** данными либо **символьными**.

Числовые данные, в свою *очередь*, могут быть дискретными и непрерывными.

Дискретные данные являются значениями признака, общее число которых конечно либо бесконечно, но может быть подсчитано при помощи натуральных чисел от одного до бесконечности.

Пример дискретных *данных*. Продолжительность маршрута троллейбуса (количество вариантов продолжительности конечно): 10, 15, 25 мин.

Непрерывные данные - данные, значения которых могут принимать какое угодно *значение* в некотором интервале. Измерение непрерывных *данных* предполагает большую *точность*.

Пример непрерывных *данных*: температура, *высота, вес, длина* и т.д.

Шкалы

Существует пять типов *шкал* измерений: номинальная, порядковая, интервальная, относительная и дихотомическая.

Номинальная шкала (nominal scale) - шкала, содержащая только категории; данные в ней не могут упорядочиваться, с ними не могут быть произведены никакие арифметические действия.

Номинальная шкала состоит из названий, категорий, имен для классификации и сортировки *объектов* или наблюдений по некоторому признаку.

Пример такой *шкалы*: профессии, город проживания, семейное положение.

Для этой *шкалы* применимы только такие операции: равно (=), не равно ($\neq$).

Порядковая шкала (ordinal scale) - шкала, в которой числа присваивают объектам для обозначения относительной позиции объектов, но не величины различий между ними.

Шкала измерений дает возможность ранжировать значения переменных. Измерения же в порядковой шкале содержат информацию только о порядке следования величин, но не позволяют сказать "насколько одна величина больше другой", или "насколько она меньше другой".

Пример такой *шкалы*: место (1, 2, 3-е), которое команда получила на соревнованиях, номер студента в рейтинге успеваемости (1-й, 23-й, и т.д.), при этом неизвестно, насколько один студент успешней другого, известен лишь его номер в рейтинге.

Для этой *шкалы* применимы только такие операции: равно (=), не равно ($\neq$), больше (>), меньше (<).

Интервальная шкала (interval scale) - шкала, разности между значениями которой могут быть вычислены, однако их отношения не имеют смысла.

Эта *шкала* позволяет находить разницу между двумя величинами, обладает свойствами номинальной и порядковой шкал, а также позволяет определить количественное изменение признака.

Пример такой *шкалы*: температура воды в море утром - 19 градусов, вечером - 24, т.е. вечерняя на 5 градусов выше, но нельзя сказать, что она в 1,26 раз выше.

Номинальная и порядковая шкалы являются дискретными, а интервальная шкала - непрерывной, она позволяет осуществлять точные измерения признака и производить арифметические операции сложения, вычитания, умножения, деления.

Для этой *шкалы* применимы только такие операции: равно (=), не равно ($\neq$), больше (>), меньше (<), операции сложения (+) и вычитания (-).

Относительная шкала (ratio scale) - шкала, в которой есть определенная точка отсчета и возможны отношения между значениями шкалы.

Пример такой *шкалы*: вес новорожденного ребенка (4 кг и 3 кг). Первый в 1,33 раза тяжелее.

Цена на картофель в супермаркете выше в 1,2 раза, чем цена на базаре.

Относительные и интервальные шкалы являются числовыми.

Для этой шкалы применимы только такие операции: равно (=), не равно ($\neq$), больше (>), меньше (<), операции сложения (+) и вычитания (-), умножения (*) и деления (/).

Дихотомическая шкала (dichotomous scale) - шкала, содержащая только две категории.

Пример такой *шкалы*: пол (мужской и женский).

Пример использования разных *шкал* для измерений свойств различных *объектов*, приведен в таблице *данных*, изображенной в таблице 2.2.

Таблица 2.2. Множество измерений свойств различных объектов

Номер объекта	Профессия (номинальная шкала)	Средний балл (интервальная шкала)	Образование (порядковая шкала)
1	слесарь	22	среднее
2	ученый	55	высшее
3	учитель	47	высшее

Пример использования различных *шкал* для измерений свойств одной системы, в данном случае температурных условий, приведен в таблице *данных*, изображенной в таблице 2.3.

Таблица 2.3. Множество измерений свойств одной системы

Дата измерения	Облачность (номинальная	Температура в 8 часов утра (интервальная	Сила ветра (порядковая

	шкала)	шкала)	шкала)
1 сентября	облачно	22 deg C	Ветер сильный
2 сентября	пасмурно	17 deg C	Ветер слабый
3 сентября	ясно	23 deg C	Ветер очень сильный

Выводы. В этой части лекции мы рассмотрели понятие *данных*, *объекта* и *атрибута*, их характеристики.

Также мы обсудили типы *шкал*. Номинальная шкала описывает *объекты* или наблюдения в терминах качественных признаков. На один шаг далее идут порядковые шкалы, позволяющие упорядочивать наблюдения или *объекты* по определенной характеристике. Интервальные и относительные шкалы более сложны, в них возможно определение количественного значения признака.

Типы наборов данных

Данные, состоящие из записей

Наиболее часто встречающиеся *данные* - данные, состоящие из записей (record data) [7]. Примеры таких *наборов данных*: табличные данные, матричные данные, документальные данные, транзакционные или операционные.

Табличные данные - данные, состоящие из записей, каждая из которых состоит из фиксированного набора *атрибутов*.

Транзакционные данные представляют собой особый тип *данных*, где каждая запись, являющаяся транзакцией, включает набор значений.

Пример транзакционной базы *данных*, содержащей перечень покупок клиентов магазина, приведен на рис. 2.1.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Рис. 2.1. Пример транзакционных данных

Графические данные

Примеры графических *данных*: WWW-данные; молекулярные структуры; графы (рис. 2.2); карты.

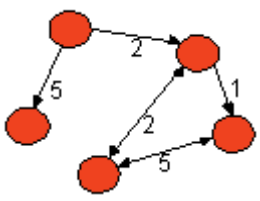

Рис. 2.2. Пример графа

С помощью карт, например, можно отследить изменения *объектов* во времени и пространстве, определить характер их распределения на плоскости или в пространстве. Преимуществом графического представления *данных* является большая простота их восприятия, чем, например, табличных *данных*.

Пример карты, являющейся картой Кохонена (моделью нейронных сетей, которые будут рассмотрены в одной из лекций нашего курса), представлен на рис. 2.3.

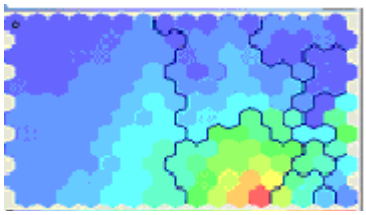

Рис. 2.3. Пример данных типа "Карта Кохонена"

Химические данные

Химические данные представляют собой особый тип *данных*. Пример таких *данных*: Benzene Molecule: C_6H_6 (рис. 2.4)

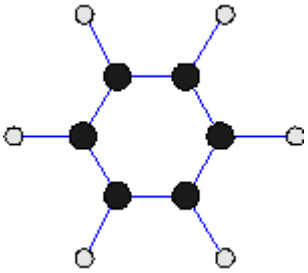

Рис. 2.4. Пример химических данных

Согласно опросу на сайте Kdnuggets, www.kdnuggets.com (апрель, 2004 г.) **"Типы анализируемых данных"**, наибольшее число опрошенных анализирует данные из "плоских" (flat table) и реляционных таблиц (26% и 24% соответственно), далее идут временные ряды (14%) и данные в виде текста (11%).

Остальные анализируемые типы *данных* в порядке убывания: web-контенты, XML, графика, аудио, видео и др.

Здесь и в следующих лекциях приводятся результаты опросов, проведенных на сайте Kdnuggets, который признан одним из наиболее авторитетных и известных сайтов в сфере Data Mining.

Форматы хранения данных

Одна из основных особенностей *данных* современного мира состоит в том, что их становится очень много. Возможны четыре аспекта работы с данными: *определение данных*, *вычисление*, манипулирование и обработка (сбор, передача и др.).

При манипулировании данными используется структура *данных* типа "*файл*". Файлы могут иметь различные форматы.

Как уже было отмечено ранее, большинство инструментов *Data* Mining позволяют импортировать *данные* из различных источников, а также экспортировать результирующие данные в различные форматы.

Данные для экспериментов удобно хранить в каком-то одном формате.

В некоторых инструментах *Data* Mining эти процедуры называются импорт/экспорт *данных*, другие позволяют напрямую открывать различные источники *данных* и сохранять результаты *Data* Mining в одном из предложенных форматов.

Наиболее распространенные форматы, согласно опросу "Форматы хранения *данных* ", представлены на рис. 2.5.

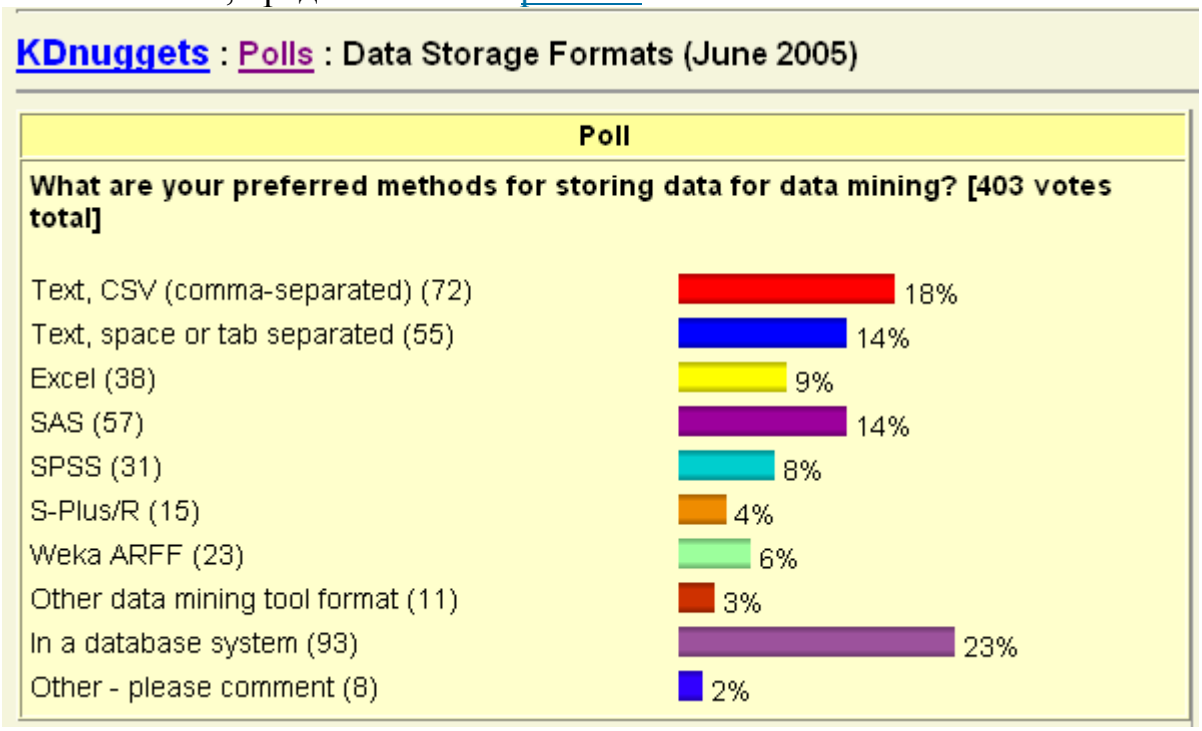

Рис. 2.5. Наиболее распространенные форматы хранения данных

Наибольшее число опрошенных (23%) предпочитают хранить данные в формате той базы *данных*, которую они используют. В формате *Text*, *CSV* - 18%, *no* 14% опрошенных хранят *данные* в формате *Text*, *space* or *tab separated* и *SAS*; в формате *Excel* - 9%, SPSS - 8%, S-Plus/R - 4%, Weka ARFF - 6%, в других форматах инструментов *Data* Mining - 2%.

Как видим из результатов опроса, наиболее распространенным форматом хранения *данных* для *Data* Mining выступают базы *данных*.

Базы данных. Основные положения

Для понимания организации *данных* в базе *данных* необходимо *знание* основных положений теории баз *данных*. Рассмотрим некоторые положения этой теории.

База данных (Database) - это особым образом организованные и хранимые в электронном виде данные.

Особым образом организованные означает, что *данные* организованы неким конкретным способом, способным облегчить их *поиск* и *доступ* к ним

для одного или нескольких приложений. Также такая организация *данных* предусматривает наличие минимальной избыточности *данных*.

Базы *данных* являются одной из разновидностей информационных технологий, а также формой хранения *данных*.

Целью создания баз *данных* является построение такой системы *данных*, которая бы не зависела от программного обеспечения, применяемых технических средств и физического расположения *данных* в ЭВМ. Построение такой системы *данных* должно обеспечивать непротиворечивую и целостную информацию. При проектировании базы *данных* предполагается многоцелевое ее использование.

База *данных* в простейшем случае представляется в виде системы двумерных таблиц.

Схема данных - описание логической структуры данных, специфицированное на языке описания данных и обрабатываемое СУБД.

Схема пользователя - зафиксированный для конкретного пользователя один вариант порядка полей таблицы.

Системы управления базами данных, СУБД

Система управления базой *данных* - это программное обеспечение, контролирующее организацию, хранение, целостность, внесение изменений, чтение и безопасность информации в базе *данных*.

СУБД (Database Management System, DBMS) представляет собой оболочку, с помощью которой при организации структуры таблиц и заполнения их данными получается та или иная база данных.

Система управления **реляционными базами данных** (Relational *Database Management* System) - это СУБД, основанная на реляционной модели *данных*.

В реляционной модели *данных* любое представление *данных* сводится к совокупности реляционных таблиц (двумерных таблиц особого типа). Системы управления реляционными базами *данных* используются для построения хранилищ *данных*.

СУБД имеет программные, технические и организационные составляющие.

Программные средства включают систему управления, обеспечивающую ввод-вывод, обработку и хранение информации, создание, модификацию и тестирование базы *данных*. Внутренними языками программирования СУБД являются языки четвертого поколения (C, C++, Pascal, *Object Pascal*). С помощью языков БД создаются приложения, базы *данных* и интерфейс пользователя, включающий экранные формы, меню, отчеты.

Аналитику при необходимости работы с конкретной СУБД, в частности, при экспорте *данных* в среду инструмента Data Mining, следует изучить особенности этой СУБД. Так, например, в базе *данных* СУБД FoxPro все таблицы и представления базы *данных* физически хранятся в отдельных

файлах, которые объединяются в одном проекте. В СУБД Access все таблицы базы *данных* хранятся в одном файле.

Для работы с конкретной базой *данных*, в том числе с целью анализа, аналитику желательно знать описание всех таблиц и их структур (*атрибутов*, типов *данных*), количество записей в таблице, а также связи между таблицами. Иногда для этих целей используется словарь *данных*.

К базам *данных*, а также к СУБД предъявляются такие требования:

1. высокое быстродействие;
2. простота обновления *данных* ;
3. независимость *данных* ;
4. возможность многопользовательского использования *данных* ;
5. безопасность *данных* ;
6. стандартизация построения и эксплуатации БД (фактически СУБД);
7. адекватность отображения *данных* соответствующей предметной области;
8. дружелюбный интерфейс пользователя.

Высокое быстродействие предусматривает малое время отклика, т.е. малый промежуток времени от момента запроса к базе *данных* до момента реального получения *данных*.

Независимость данных - это возможность изменения логической и физической структуры базы *данных* без изменения представлений пользователей.

Независимость *данных* обеспечивает минимальные изменения структуры базы *данных* при изменениях стратегии доступа к данным и структуры самих исходных *данных*. Эти изменения должны быть предусмотрены на этапах концептуального и *логического проектирования* базы *данных* с обеспечением минимальных изменений на этапе физического ее проектирования.

Безопасность данных - это защита *данных* от преднамеренного или непреднамеренного нарушения секретности, искажения или разрушения. Безопасность включает два компонента: целостность и защиту *данных* от несанкционированного доступа.

Целостность данных - устойчивость хранимых *данных* к разрушению и уничтожению, связанным с неисправностями технических средств, системными ошибками и ошибочными действиями пользователей.

Целостность *данных* - точность и валидность *данных*. Целостность *данных* предполагает: отсутствие неточно введенных *данных*, защиту от ошибок при обновлении баз *данных* ; невозможность удаления (или *каскадное удаление*) связанных *данных* разных таблиц; сохранность *данных* при сбоях техники (возможность восстановления *данных*) и др.

Защита данных от несанкционированного доступа предполагает ограничение доступа к определенным данным базы и достигается введением мер безопасности: разграничение прав доступа к данным различных

пользователей в зависимости от выполняемых ими функций и/или должностных обязанностей; введением защиты в виде паролей; использованием представлений, т.е. таблиц, которые являются производными от исходных и предназначены для работы конкретных пользователей для решения конкретных задач.

Стандартизация обеспечивает преемственность поколений конкретной СУБД, упрощает взаимодействие баз *данных* одного поколения СУБД с одинаковыми и различными моделями *данных*.

СУБД отвечает за обработку запросов к базе *данных* и получение ответа. Способы хранения *данных* могут быть различными: модель *данных* может быть как реляционной, так и многомерной, сетевой или иерархической.

Классификация видов данных

Какими могут быть данные? Ниже приведено несколько классификаций.

Реляционные данные - это данные из реляционных баз (таблиц).

Многомерные данные - это данные, представленные в кубах *OLAP*.

Измерение (dimension) или ось - в многомерных *данных* - это собрание *данных* одного и того же типа, что позволяет структурировать многомерную базу *данных*.

По критерию постоянства своих значений в ходе решения задачи данные могут быть:

- переменными;
- постоянными;
- условно-постоянными.

Переменные данные - это такие данные, которые изменяют свои значения в процессе решения задачи.

Постоянные данные - это такие данные, которые сохраняют свои значения в процессе решения задачи (математические *константы, координаты* неподвижных *объектов*) и не зависят от внешних факторов.

Условно-постоянные данные - это такие данные, которые могут иногда изменять свои значения, но эти изменения не зависят от процесса решения задачи, а определяются внешними факторами.

Данные, в зависимости от тех функций, которые они выполняют, могут быть **справочными, оперативными, архивными.**

Следует различать данные за период и точечные данные. Эти различия важны при проектировании системы сбора информации, а также в процессе измерений.

- данные за период;
- точечные данные.

Данные за период характеризуют некоторый период времени. Примером *данных* за период могут быть: *прибыль* предприятия за месяц, средняя температура за месяц.

Точечные данные представляют *значение* некоторой переменной в конкретный момент времени. Пример точечных *данных*: *остаток* на счете на первое число месяца, температура в восемь часов утра.

Данные бывают первичными и вторичными. **Вторичные данные** - это данные, которые являются результатом определенных вычислений, примененных к **первичным данным**. Вторичные данные, как правило, приводят к ускоренному получению ответа на *запрос* пользователя за счет увеличения объема хранимой информации.

Метаданные

В завершение лекции о *данных* рассмотрим понятие метаданных.

Метаданные (Metadata) - это данные о *данных*.

В состав метаданных могут входить: каталоги, справочники, реестры.

Метаданные содержат сведения о составе *данных*, содержании, статусе, происхождении, местонахождении, качестве, форматах и формах представления, условиях доступа, приобретения и использования, авторских, имущественных и смежных с ними правах на данные и др.

Метаданные - важное понятие в управлении хранилищем *данных*.

Метаданные, применяемые при управлении хранилищем, содержат информацию, необходимую для его настройки и использования. Различают бизнес-*метаданные* и оперативные *метаданные*.

Бизнес-метаданные содержат бизнес-термины и определения, принадлежность *данных* и правила оплаты услуг хранилища.

Оперативные метаданные - это *информация*, собранная во время работы хранилища *данных*:

- происхождение перенесенных и преобразованных *данных* ;
- статус использования *данных* (активные, архивированные или удаленные);
- данные мониторинга, такие как статистика использования, сообщения об ошибках и т.д.

Метаданные хранилища обычно размещаются в репозитории. Это позволяет использовать *метаданные* совместно различным инструментам, а также процессам при проектировании, установке, эксплуатации и администрировании хранилища.

Выводы. В лекции были рассмотрены понятие *данных*, *объектов* и *атрибутов*, их характеристики, типы *шкал*, понятие *набора данных* и его типы. Описаны возможные форматы хранения *данных*. Введены понятия базы *данных*, системы управления базами *данных*, метаданных.