Структурированные данные [M.004]

Данные, описывающие реальные объекты, процессы и явления, могут быть представлены в различных формах и иметь разные тип и вид.

Формы представления данных

Определение

> *Данные – сведения, характеризующие систему, явление, процесс или объект, представленные в определенной форме и предназначенные для дальнейшего использования.*

По степени структурированности выделяют следующие формы представления данных:

- неструктурированные;
- структурированные;
- слабоструктурированные.

К *неструктурированным* относятся данные, произвольные по форме, включающие тексты и графику, мультимедиа (видео, речь, аудио). Эта форма представления данных широко используется, например, в Интернете, а сами данные предоставляются пользователю в виде отклика поисковыми системами.

Структурированные данные отражают отдельные факты предметной области. Структурированными называются данные, определенным образом упорядоченные и организованные с целью обеспечения возможности применения к ним некоторых действий (например, визуального или машинного анализа). Это основная форма представления сведений в базах данных.

Организация того или иного вида хранения данных (структурированных или неструктурированных) связана с обеспечением доступа к ним. Под доступом понимается возможность выделения элемента данных (или множества элементов) среди других элементов по каким-либо признакам с целью выполнения некоторых действий над элементом.

Одной из самых распространенных моделей хранения структурированных данных является **таблица**. В ней все данные упорядочиваются в двумерную структуру, состоящую из столбцов и строк (рисунок 1).

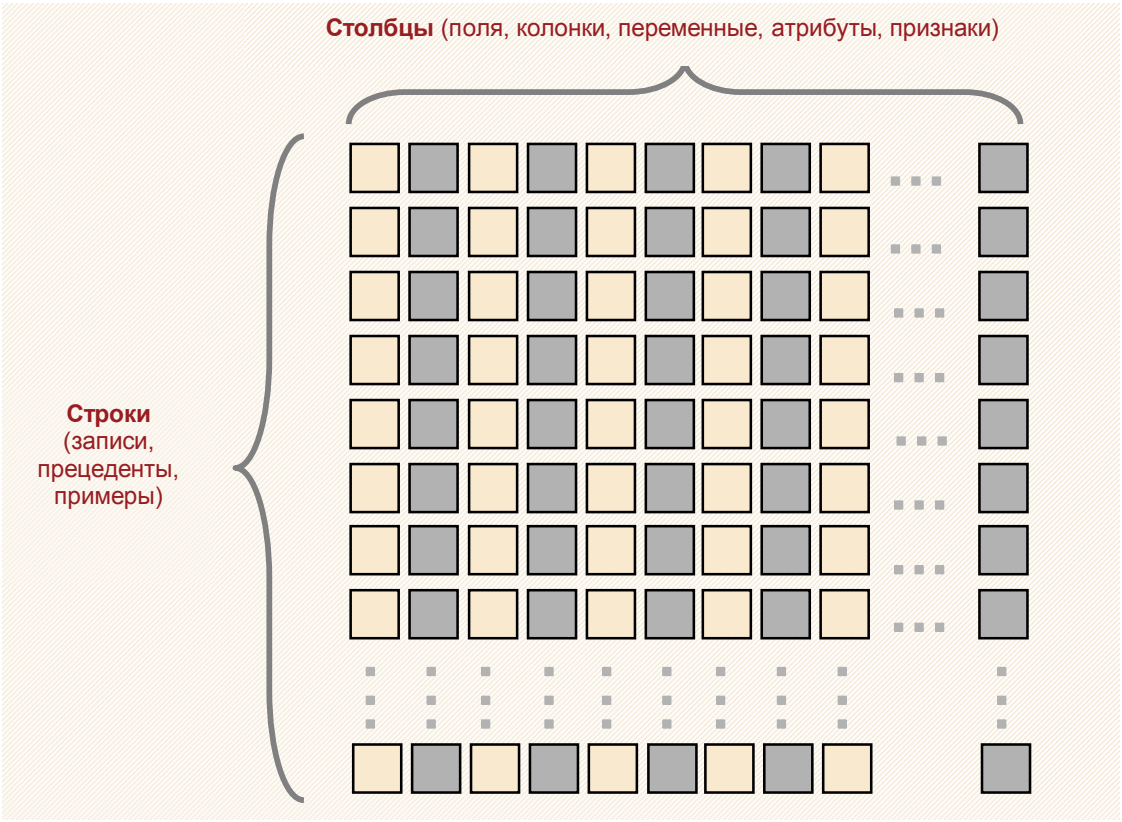

Рисунок 1 – Структурированный набор данных

В ячейках такой таблицы содержатся элементы данных: символы, числа, логические значения.

Неструктурированные данные непригодны для обработки напрямую методами анализа данных, поэтому такие данные подвергаются специальным приемам структуризации, причем сам характер данных в процессе структуризации может существенно измениться. Например, в анализе текстов (Text Mining) при структурировании из исходного текста может быть сформирована таблица с частотами встречаемости слов, и уже такой набор данных будет обрабатываться методами, пригодными для структурированных данных.

Слабоструктурированные данные — это данные, для которых определены некоторые правила и форматы, но в самом общем виде. Например, строка с адресом, строка в прайс-листе, ФИО и т. п. В отличие от неструктурированных, такие данные с меньшими усилиями преобразуются к структурированной форме, однако без процедуры преобразования они тоже непригодны для анализа. На рисунке 2 приведен пример стандартизации строки с адресом.

Рисунок 2 – Стандартизация слабоструктурированных данных

Подавляющее большинство методов анализа данных работает только с хорошо **структурированными данными**, представленными в **табличном виде**, поэтому дальнейшее

изложение во всем курсе ведется применительно к структурированным данным. Сбор информации в структурированном виде осуществляется на этапе подготовки данных к анализу и обсуждается в соответствующей теме.

Типы данных

Все структурированные данные делятся на пять типов:

- целый (количество товара, код товара и т. п.);
- вещественный (цена, скидка и т. п.);
- строковый (фамилия, наименование, адрес, пол, образование и т. п.);
- логический;
- дата/время.

Среди строковых данных можно выделить два подтипа: **упорядоченные (ординальные)** и **категориальные**. В обоих случаях переменная относится к одному из значений дискретного набора классов $c_1,\ldots,c_k$ и описывает некоторые качественные свойства объекта. Но если в случае ординальных данных эти классы можно упорядочить, то в случае категориальных данных — нельзя. При сравнении категориальных данных применимы только две операции: «равно» (=) и «не равно» ($\neq$). В таблице 1 ниже только одно поле Образование является упорядоченным, а все остальные — категориальными.

Таблица 1 — Множество данных одной системы

Код заявки	Фамилия	Образование	Профессия	Город
Z-01	Иванов	высшее	инженер	Москва
Z-02	Кузнецова	среднее	бухгалтер	Коломна

Значения в столбце Образование можно упорядочить по убыванию: высшее > среднее > начальное > …

Со значениями полей Код заявки, Фамилия, Профессия, Город проделывать упорядочивание бессмысленно, применимы лишь операции «равно» и «не равно».

Виды данных

По виду данные делятся на *непрерывные* и *дискретные*.

Определение

> *Непрерывные данные — данные, значения которых могут принимать какое угодно значение в некотором интервале. Над непрерывными данными можно производить арифметические операции сложения, вычитания, умножения, деления, и они имеют смысл.*

Примерами непрерывных данных являются возраст, любые стоимостные показатели, количественные оценки (количество товара, объем отгрузки, вес отгрузки).

Определение

> *Дискретные данные — значения признака, общее число которых конечно либо бесконечно, но может быть подсчитано при помощи натуральных чисел от одного до бесконечности. С дискретными данными не могут быть произведены никакие арифметические действия, либо они не имеют смысла.*

Дискретными данными являются все данные строкового и логического типа. Дискретными могут быть и числовые данные. Например, поле Код товара, принимающее значения целого типа (как правило), дискретно, так как операции сложения, вычитания, умножения над Код товара не имеют смысла. Соответствие возможных видов данных типам данных приведено в таблице 2.

Таблица 2 — Соответствие между типами и видами данных

Тип данных	Вид данных	
	Непрерывный	Дискретный
Целый	+	+
Вещественный	+	+
Строковый		+
Логический		+
Дата/время	+	+

Аналитику важно понимать природу данных для выбора адекватных методов их предобработки, очистки и построения моделей.

Представления наборов данных

По отношению к задаче анализа наборы данных могут быть **упорядоченными** и **неупорядоченными**.

В *упорядоченном* наборе данных каждому столбцу соответствует один фактор, а в каждую строку заносятся упорядоченные по какому-либо признаку события с интервалом периода между строками. Часто таким признаком выступает время. На рисунке 3 приведены примеры упорядоченных наборов данных — временной ряд (слева, упорядочен по дате) и ряд показаний датчика зонда (справа, упорядочен по глубине скважины).

Дата	Количество	Сумма
01.01.2004	4	283.31
01.01.2004	1	72.48
01.01.2004	1	173.32
02.01.2004	6	294.84
02.01.2004	2	405.76
02.01.2004	12	303.13
02.01.2004	1	210.5
03.01.2004	6	521.16
03.01.2004	3	156.96

Глубина	BK	DS
887.9	8.85	0.218
888.1	9.627	0.216
888.3	14.584	0.217
888.5	21.647	0.215
888.7	17.172	0.216
888.9	6.118	0.215
889.1	2.886	0.217
889.3	2.506	0.219

Рисунок 3 – Примеры упорядоченных наборов данных

В *неупорядоченном* наборе каждому столбцу соответствует фактор, а в каждую строку заносится пример (ситуация, прецедент), соответственно, упорядоченность строк не требуется. Пример такого набора данных приведен на рисунке 4.

Номер	Банк	Реутеры	Филиалы	Город	Собственные активы
2	Внешторгбанк	-	32	Москва	23236327
3	Газпромбанк	GZPM	27	Москва	9255041
4	ООО "Международный Промышленный банк"	TIBP	4	Москва	26409116
5	Международный Московский Банк	IMBX	1	Москва	1176462
6	ОАО "АЛЬФА-БАНК"	ALFM	17	Москва	12446938
7	ОАО "ПСБ"	ICSP	44	Санкт-Петербург	1275859
8	Банк Москвы	-	34	Москва	3335734
9	АКБ "РОСБАНК" (ОАО)	-	13	Москва	4691449
10	АКБ "ДИБ"	DIBM	0	Москва	2616993

Рисунок 4 – Пример неупорядоченного набора данных

Особо выделяют **транзакционные данные**. Под *транзакцией* подразумеваются несколько объектов или действий, являющихся логически связанной единицей (рисунок 5). Очень часто данный механизм используется для анализа покупок (чеков) в супермаркетах. Но в общем случае речь может идти о любых связанных объектах или действиях.

Рисунок 5 – Транзакционные данных