КЛАССИФИКАЦИЯ И КОДИРОВАНИЕ ИНФОРМАЦИИ

СИСТЕМА КЛАССИФИКАЦИИ

Классификация — система распределения объектов (предметов, явлений, процессов, понятий) по классам в соответствии с определенным признаком.

Под ***объектом*** понимается любой предмет, процесс, явление материального или нематериального свойства.

Система классификации позволяет сгруппировать объекты выделить определенные классы, которые будут характеризоваться рядом общих свойств.

Классификация объектов — это **процедура группировки на качественном уровне, направленная на выделение однородных свойств**. Применительно к информации как к объекту классификации выделенные классы называют ***информационными объектами***.

Свойства информационного объекта определяются информационными параметрами, называемыми ***реквизитами***.

Реквизиты представляются либо **числовыми данными**, например вес, стоимость, год, либо **признаками**, например цвет, марка машины, фамилия.

Реквизит — логически неделимый информационный элемент, описывающий определенное свойство объекта, процесса, явления и т.п.

При любой классификации желательно, чтобы соблюдались следующие требования:

- **полнота охвата объектов рассматриваемой области;**
- **однозначность реквизитов;**
- **возможность включения новых объектов.**

Классификатор — систематизированный свод наименований и кодов классификационных группировок.

При классификации широко используются понятия ***классификационный признак*** и ***значение классификационного признака***, которые позволяют установить сходство или различие объектов. Возможен подход к классификации с объединением этих двух понятий в одно; названное как признак классификации.

Разработаны три метода классификации объектов:

- **иерархический,**
- **фасетный,**
- **дескрипторный.**

Эти методы различаются разной стратегией применения классификационных признаков.

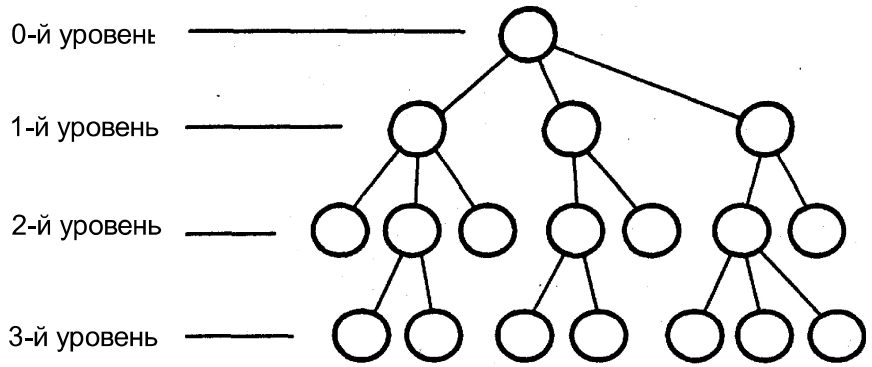
Иерархическая система классификации

Иерархическая система классификации **(рис.) строится следующим образом:**

- **исходное множество элементов составляет 0-й уровень и делится в зависимости от выбранного классификационного признака на классы (группировки), которые образуют 1-й уровень;**

- **каждый класс 1-го уровня в соответствии со своим, характерным для него классификационным признаком делится на подклассы, которые образуют 2-й уровень;**

- **каждый класс 2-го уровня аналогично делится на группы, которые образуют 3-й уровень, и т.д.**



Учитывая достаточно жесткую процедуру построения структуры классификации, необходимо перед началом работы определить ее цель, т.е. какими свойствами должны обладать объединяемые в классы объекты.

Эти свойства принимаются в дальнейшем за признаки классификации.

В иерархической системе классификации каждый объект на любом уровне должен быть отнесен к одному классу, который характеризуется конкретным значением выбранного классификационного признака.

Для последующей группировки в каждом новом классе необходимо задать свои классификационные признаки и их значения.

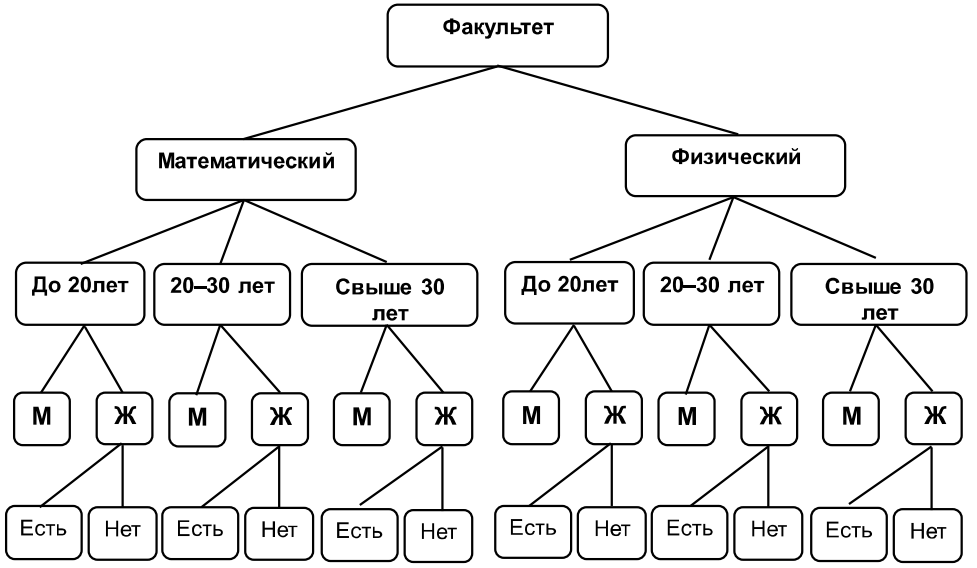
Количество уровней классификации, соответствующее числу признаков, выбранных в качестве основания деления, характеризует ***глубину классификации.***

Достоинства иерархической системы классификации**:**

- **простота построения;**
- **использование независимых классификационных признаков в различных ветвях иерархической структуры.**

Недостатки иерархической системы классификации:

- **жесткая структура, которая приводит к сложности внесения изменений,** так как приходится перераспределять все классификационные группировки;
- **невозможность группировать объекты по заранее не предусмотренным сочетаниям признаков.**



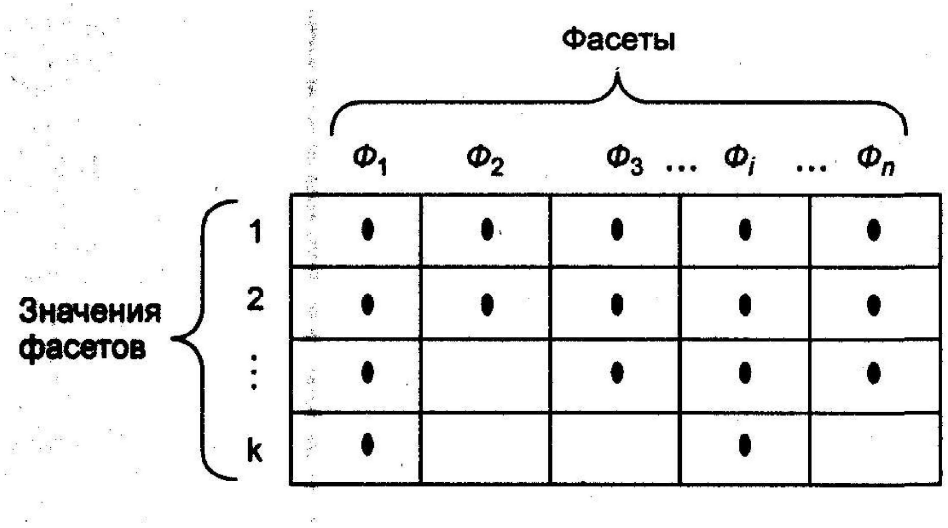
Фасетная система классификации

Фасетная система классификации в отличие от иерархической **позволяет выбирать признаки классификации независимо как друг от друга, так и от семантического содержания классифицируемого объекта**.

Признаки классификации называются ***фасетами*** (facet — рамка).

Каждый фасет *(Фi)* содержит совокупность однородных значений данного классификационного признака. Причем **значения в фасете могут располагаться в произвольном порядке**, хотя предпочтительнее их упорядочение.

Схема построения фасетной системы классификации в виде таблицы отображена на рисунке.



Названия столбцов соответствуют выделенным классификационным признакам (фасетам), обозначенным *Ф1,* Ф2, ..., Фi, …, Фn. Например, цвет, размер одежды, вес и т.д.

Произведена нумерация строк таблицы. В каждой клетке таблицы хранится конкретное значение фасета. Например, фасет *цвет,* обозначенный Ф2, содержит значения: красный, белый, зеленый, черный, желтый.

Процедура классификации состоит в присвоении каждому объекту соответствующих значений из фасетов. Для каждого объекта задается конкретная группировка фасетов структурной формулой, в которой отражается их порядок следования:

$$K_S = (\Phi_1, \Phi_2, \ldots, \Phi_i, \ldots, \Phi_n)$$

где *Фi — i-й* фасет;
n — количество фасетов.

При построении фасетной системы классификации необходимо, чтобы значения, используемые в различных фасетах, не повторялись. **Фасетную систему легко можно модифицировать, внося изменения в конкретные значения любого фасета.**

Достоинства фасетной системы классификации:
• **возможность создания большой емкости классификации,** т.е. использования большого числа признаков классификации и их значений для создания группировок,
• **возможность простой модификации всей системы классификации без изменения структуры существующих группировок.**

Недостатком фасетной системы классификации является **сложность ее построения, так как необходимо учитывать все многообразие классификационных признаков**.

Ks = (Факультет, Возраст, Пол, Дети)

Название факультета	**Возраст**	**Пол**	**Дети**
Радиотехнический	До 20 лет	М	Есть
Машиностроительный	20—30 лет	Ж	Нет
Коммерческий	Свыше 30 лет		
Информационные системы			
Математический			

Дескрипторная система классификации

Для организации поиска информации, для ведения тезаурусов (словарей) **эффективно используется дескрипторная** (описательная) **система классификации**, **язык которой приближается к естественному языку описания информационных объектов**. Особенно широко она используется в библиотечной системе поиска.

Суть дескрипторного метода классификации заключается в следующем:

• **отбирается совокупность ключевых слов или словосочетаний**, описывающих определенную предметную область или совокупность однородных объектов;
• выбранные ключевые слова и словосочетания подвергаются ***нормализации***, т.е. из совокупности синонимов выбирается один или несколько наиболее употребимых;
• создается ***словарь дескрипторов.***

Пример.
В качестве предметной области выбирается учебная деятельность в высшем учебном заведении. ***Ключевыми словами*** могут быть выбраны: **студент, обучаемый, учащийся, преподаватель, учитель, педагог, лектор, ассистент, доцент, профессор, коллега, факультет, подразделение университета, аудитория, комната, лекция, практическое занятие, занятие** и т.д.
Среди указанных ключевых слов встречаются синонимы, например: **студент, обучаемый, учащийся; преподаватель, учитель, педагог**; факультет, подразделение университета и т.д.
После нормализации словарь дескрипторов будет состоять из следующих слов: **студент, преподаватель, лектор, ассистент, доцент, профессор, факультет, аудитория, лекция, практическое занятие** и т.д.

Между дескрипторами устанавливаются связи, которые позволяют расширить область поиска информации.

Связи могут быть трех видов:
• ***синонимические***, указывающие некоторую совокупность ключевых слов как синонимы;
• ***родо-видовые***, отражающие включение некоторого класса объектов в более представительный класс;
• ***ассоциативные***, соединяющие дескрипторы, обладающие общими свойствами.

Пример.
Синонимическая связь: студент—учащийся—обучаемый.
Родо-видовая связь: университет — факультет — кафедра.
Ассоциативная связь: студент — экзамен — профессор — аудитория.

СИСТЕМА КОДИРОВАНИЯ

Система кодирования применяется для замены названия объекта на условное обозначение (код) в целях обеспечения удобной и более эффективной обработки информации.

Система кодирования **- совокупность правил кодового обозначения объектов.**

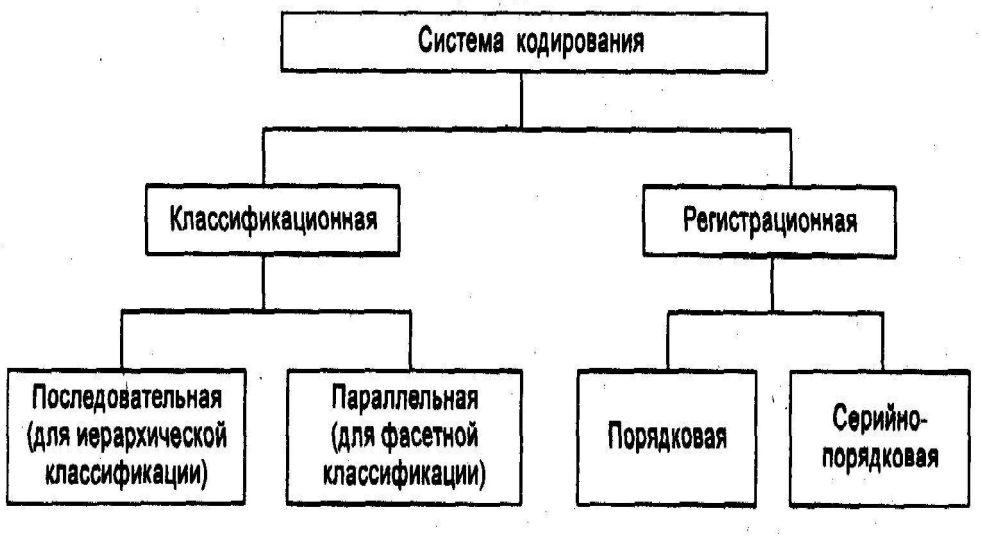
Код строится на базе алфавита, состоящего из букв, цифр и других символов.
Код характеризуется:
• ***длиной*** **—** число позиций в коде;
• ***структурой*** — порядок расположения в коде символов, используемых для обозначения классификационного признака.

Процедура присвоения объекту кодового обозначения называется *кодированием.*

Можно выделить две группы методов, используемых в системе кодирования (рис.), которые образуют:

• ***классификационную систему кодирования*** ориентированную на проведение предварительной классификации объектов либо на основе иерархической системы, либо на основе фасетной системы;

• ***регистрационную систему кодирования***, не требующую предварительной классификации объектов.



Классификационное кодирование

Классификационное кодирование применяется после проведения классификации объектов. Различают ***последовательное*** и ***параллельное*** **кодирование**.

Последовательное **кодирование используется для** ***иерархической*** **классификационной структуры**.
Суть метода заключается в следующем: сначала записывается код старшей группировки 1-го уровня, затем код группировки 2-го уровня, затем код группировки 3-го уровня и т.д. **В результате получается кодовая комбинация, каждый разряд которой содержит информацию о специфике выделенной группы на каждом уровне иерархической структуры.**

Последовательная система кодирования обладает теми же достоинствами и недостатками, что и иерархическая система классификации.

Параллельное **кодирование используется для фасетной системы классификации**.

Суть метода заключается в следующем: все фасеты кодируются независимо друг от друга; для значений каждого фасета выделяется определенное количество разрядов кода.

Параллельная система кодирования обладает теми же достоинствами и недостатками, что и фасетная система классификации.

Регистрационное кодирование

Регистрационное кодирование используется для однозначной идентификации объектов и не требует предварительной классификации объектов. Различают ***порядковую*** и ***серийно-порядковую*** систему.

Порядковая **система кодирования предполагает последовательную нумерацию объектов числами натурального ряда.** Этот порядок может быть случайным или определяться после предварительного упорядочения объектов, например по алфавиту.

Этот метод применяется в том случае, когда количество объектов невелико, например кодирование названий факультетов университета, кодирование студентов в учебной группе.

Серийно-порядковая **система кодирования предусматривает предварительное выделение групп объектов, которые составляют серию**, а **затем в каждой серии производится порядковая нумерация объектов.** Каждая серия также будет иметь порядковую нумерацию. По своей сути серийно-порядковая система является смешанной: классифицирующей и идентифицирующей. Применяется тогда, когда количество групп невелико.

Пример. Все студенты одного факультета разбиваются на учебные группы (в данной терминологии — серии), для которых используется порядковая нумерация. Внутри каждой группы производится упорядочение фамилий студентов по алфавиту и каждому студенту присваивается номер.

Классификационное кодирование