Сигнальный экземпляр

Камчатский государственный технический университет

Кафедра информационных систем

Е.А. Малова

ТЕОРИЯ ЭКОНОМИЧЕСКИХ ИНФОРМАЦИОННЫХ СИСТЕМ

Конспект лекций для студентов специальности 351400 «Прикладная информатика (в экономике)» очной и заочной форм обучения

Петропавловск-Камчатский

2005

УДК 338.46:002.52/54
ББК 32.973.202
М19

Рецензент

С.В. Чебанюк,
доцент кафедры информационных систем КамчатГТУ

Малова Е.А.

М19 Теория экономических информационных систем. Конспект лекций для студентов специальности 351400 «Прикладная информатика (в экономике)» очной и заочной форм обучения. – Петропавловск-Камчатский: КамчатГТУ, 2005. – 39 с.

Конспект лекций составлен в соответствии с требованиями к обязательному минимуму содержания основной образовательной программы подготовки специалиста по специальности 351400 «Прикладная информатика (в экономике)» государственного образовательного стандарта высшего профессионального образования.

Рекомендовано к изданию решением президиума учебно-методического совета КамчатГТУ (протокол № 4 от 24 апреля 2005 г.).

УДК 338.46:002.52/54
ББК 32.973.202

Камчатский государственный технический университет

Кафедра информационных систем

Е.А. Малова

ТЕОРИЯ ЭКОНОМИЧЕСКИХ ИНФОРМАЦИОННЫХ СИСТЕМ

Конспект лекций для студентов специальности 351400 «Прикладная информатика (в экономике)» очной и заочной форм обучения

Петропавловск-Камчатский
2005

УДК 338.46:002.52/54
ББК 32.973.202
М19

Рецензент

С.В. Чебанюк,
доцент кафедры информационных систем КамчатГТУ

Малова Е.А.

М19 Теория экономических информационных систем. Конспект лекций для студентов специальности 351400 «Прикладная информатика (в экономике)» очной и заочной форм обучения. – Петропавловск-Камчатский: КамчатГТУ, 2005. – 39 с.

Конспект лекций составлен в соответствии с требованиями к обязательному минимуму содержания основной образовательной программы подготовки специалиста по специальности 351400 «Прикладная информатика (в экономике)» государственного образовательного стандарта высшего профессионального образования.

Рекомендовано к изданию решением президиума учебно-методического совета КамчатГТУ (протокол № 4 от 24 апреля 2005 г.).

УДК 338.46:002.52/54
ББК 32.973.202

ТЕМА 1. ОБЩАЯ ХАРАКТЕРИСТИКА ИНФОРМАЦИИ И ИНФОРМАЦИОННЫХ СИСТЕМ

1. **Предмет и цели курса.**
2. **Основные понятия.**
3. **Единицы информации.**

Предметом является экономическая информация, рассматриваемая как предмет и продукт труда, на основе которой возможно улучшить процесс управления любым элементом предметной области.

Целью курса является изучение теоретических основ строения информации, ее разновидности, структурной организации данных, методов и средств описания ЭИС.

Понятие информации

Распространенным является взгляд на информацию как на ресурс, аналогичный материальным, трудовым и денежным ресурсам.

Эта точка зрения отражается в следующем определении:

«Информация – новые сведения, позволяющие улучшить процессы, связанные с преобразованием вещества, энергии и самой информации» (с точки зрения отображения информации).

По ВИННЕРУ: «информация – это обозначение содержания, полученного из внешнего мира в процессе нашего приспособления к нему».

С точки зрения потребителя: «Информация – это новые сведения, принятые, понятые и оцененные конечным потребителем как полезные». Информацией являются сведения, расширяющие запас знаний конечного потребителя.

ШЕНОН выделил 5 компонентов передачи информации (информация на пути от источника к потребителю проходит через ряд преобразователей и декодирующих устройств, вычислительную машину, обрабатывающую информацию по определенному алгоритму).

- источник,
- прием,
- передатчик,
- пользователь,
- канал связи.

Данные – это формальное представление информации, написанное на некотором языке и перенесенное на материальные носитель; или - это набор утверждений, фактов или цифр лексически и синтаксически взаимосвязанных м/д собой.

Существует 3 основных аспекта изучения информации и данных:

1. с точки зрения смысла – семантический анализ, основным критерием является оценка новизны полученной информации, расширение знаний;

2. синтаксический анализ – проводится с точки зрения формы представления информации;

3. прагматический анализ – устанавливает ценность, полезность данных, информации.

Экономическая информация имеет ряд особенностей:

1. преобладание алфавитно-цифровой записи информации,
2. высокий удельный вес исходных данных и условной информации,
3. широкое применение документальной формы носителей информации,
4. необходимость удобства восприятия результатов обработки информации.

Эк. данные – это сведения, которые могут быть получены путем преобразования: наблюдение, регистрация, арифметические и логические преобразования; и использования для передачи и обработки в удобной для пользователя форме.

Виды информационного пространства

1. неструктурированная информация,
2. частично или слабо структурированное пространство (описание поделено на части, выделено главное),
3. структурированное (основные языковые атрибуты, синтаксис),
4. формализованное (выделены параметры оценки информации),
5. машинно-структурированное (данные введены в компьютер).

Понятие системы

Система – это совокупность взаимосвязанных объектов (элементов), функционирующих для достижения общей цели.

Для системы характерно изменение состояний объектов, которое с течением времени происходит в результате взаимодействия объектов в различных процессах и с внешней средой. В результате такого поведения системы важно соблюдение следующих принципов:

- ***эмерджентности***, т. е. целостности системы на основе общей структуры, когда поведение отдельных объектов рассматривается с позиции функционирования всей системы.
- ***гомеостазиса,*** то есть обеспечения устойчивого функционирования системы и достижения общей цели,
- ***адаптивности к изменениям*** внешней среды и управляемости посредством воздействия на элементы системы,
- ***обучаемости*** путем изменения структуры системы в соответствии с изменением целей системы.

Системы могут быть:

1. открытыми, – закрытыми,

2. искусственными (созданные человеком), – естественными,

3. статическими (собирают данные без прогнозов на развитие), – динамическими (осуществляют планирование, прогнозирование,. анализ данных).

Информационная система – это БД, концептуальная схема, информационный процессор, образующие совместно систему хранения и манипулирования данными.

БД – это набор сообщений, организованных определенным образом в виде единиц информации.

Концептуальная схема – описание структуры всех единиц информации, хранимых в БД.

Информационный процессор – это вычислительная система или СУБД, которая выполняет операции с БД и концептуальной схемой.

ЭИС – совокупность элементов и процессов обработки экономических данных, обеспечивающая потребность управляющего звена в экономической информации.

ЭИС – совокупность организационных, технических, программных и информационных средств, объединенных в единую систему с целью сбора, хранения и выдачи необходимой информации, предназначенной для выполнения функций управления. ЭИС связывает объект управления и систему управления м/д собой и с внешней средой через информационные потоки.

ИП 1 – информационный поток из внешней среды в систему управления, который, с одной стороны. Представляет поток нормативной информации, создаваемый государственными учреждениями в части законодательства, а, с другой стороны, – поток информации о конъюнктуре рынка, создаваемый конкурентами, потребителями, поставщиками;

ИП 2 – информационный поток из системы управления во внешнюю среду, а именно: отчетная информация, прежде всего финансовая информация в гос органы, инвесторам, кредиторам, потребителям; маркетинговая информация потенциальным потребителям);

ИП 3 – информационный поток из систем управления на объект управления (прямая кибернетическая связь). Представляющий совокупность плановой, нормативной и распорядительной информации для осуществления хозяйственных процессов;

ИП 4 – информационный поток от объекта управления в систему управления (обратная кибернетическая связь), который отражает учетную информацию о состоянии объекта управления экономической системой (сырья, материалов, денежных, энергетических, трудовых ресурсов, готовой продукции и выполненных услугах) в результате выполнения хозяйственных процессов.

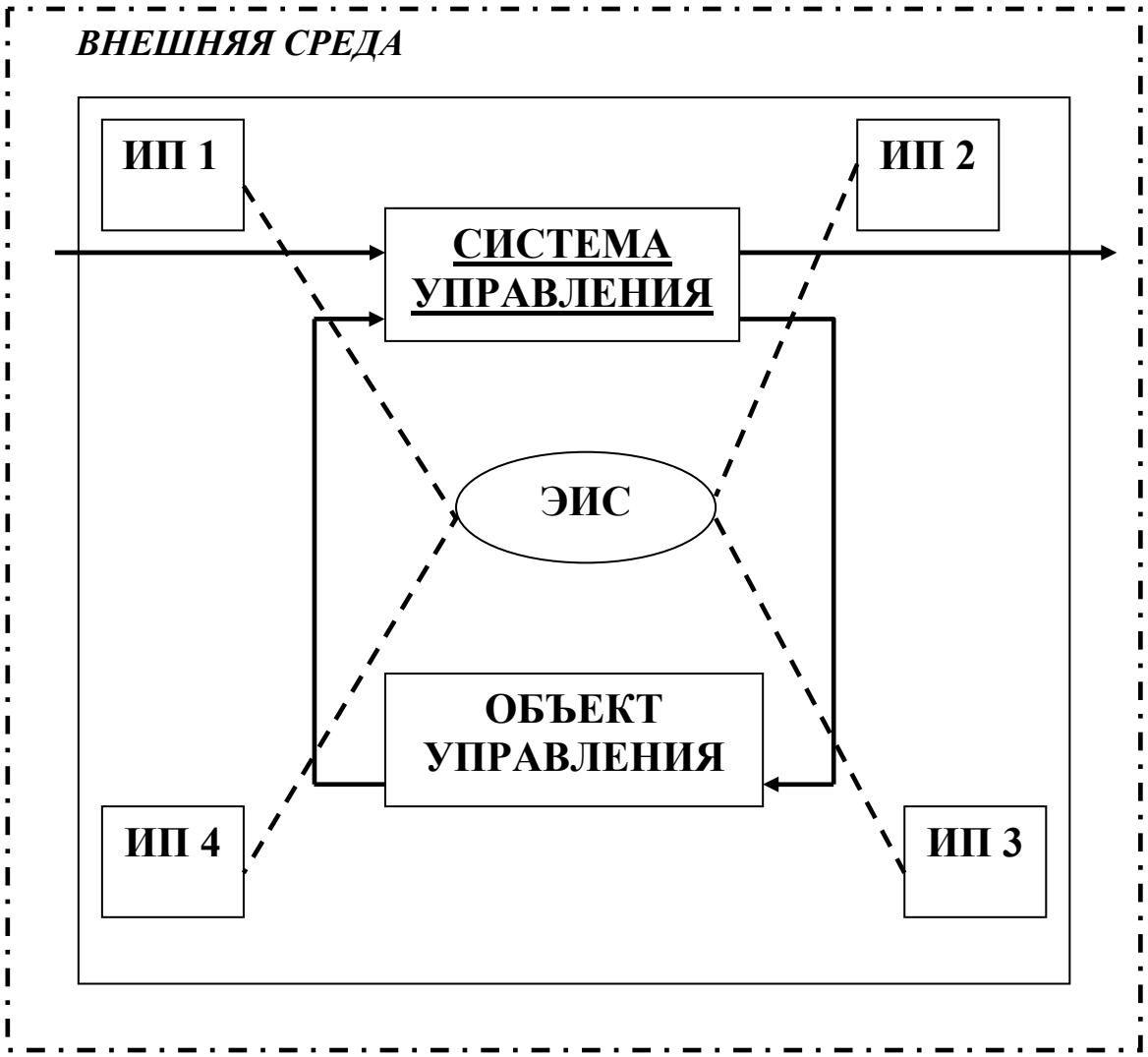


Рисунок 1. Структура экономической системы.

Для составления БД необходимы **единицы информации.**

ЕИ – это набор символов, которому придается определенный смысл.

Каждая ЕИ характеризуется следующими свойствами:

1. и м я (отражает уникальность) – уникальной наименование ЕИ в процессе ее обработки;

2. с т р у к т у р а (составные части) – реквизитный состав с учетом вхождения данной ЕИ в другую – более высокого уровня, структура реквизита является описанием его формата – множеством допустимых символов в каждой позиции;

3. з н а ч е н и е (величина, характеризующая некоторое свойство объекта, явления, процесса в конкретных обстоятельствах);

4. м е т о д о р г а н и з а ц и и з н а ч е н и й – это точное описание множества значений и их взаимосвязей, которые поддерживаются м/д элементами ЕИ.

5. о п е р а ц и и над и м е н е м, с т р у к т у р о й, з н а ч е н и я м и:

- операции над именем (присвоение, переименование, открытие-закрытие данного имени),

- =/=/= над структурой (композиция, декомпозиция);
- =/=/= над значениями (возможность логических и математич. операций над значениями).

Виды ЕИ

Существует две основные единицы информации: атрибут и составная единица информации.

1. атрибут (реквизит) – это логически неделимый элемент любой сложной информационной совокупности. **Атрибут – это информационное отображение отдельного свойства некоторого объекта, процесса или явления**

Характеризуется именем, структурой (тип описания длины и формата) и значением.

2. СЕИ представляют собой набор из атрибутов и, возможно, других СЕИ, ассоциативно связанных м/д собой некоторыми отношениями.

Характеризуются именем, структурой (представляется в виде графа), значением.

Над ними могут проводиться операции:

1. нормализация – операция перехода от СЕИ с произвольной структурой к СЕИ с двухуровневой структурой;
2. свёртка (обратный процесс);
3. композиция – объединение нескольких СЕИ в 1.
4. декомпозиция;
5. выборка – выделение подмножества значений СЕИ, которые удовлетворяют заранее поставленным условиям выборки;
6. корректировка – означает выполнение одной из операций

- добавление нового значения СЕИ,
- исключение существующего значения СЕИ,
- замена некоторого значения СЕИ на новое.

Экономические показатели

Показатель является частным случаем СЕИ.

Экономический показатель – это отношение с минимальным набором реквизита-основания и реквизитов-признаков, способное образовать осмысленный документ.

Для формирования показателей необходимо учитывать следующие закономерности:

1. если значение атрибута явл исходными данными или результатом арифметической операции, то это **основание,**
2. если значение текстовое – **это признак,**
3. если атрибут обозначает предмет или время – это **признак**,
4. если атрибут в каком-либо показателе играет роль признака или основания. То он будет играть эту же роль и в других показателях.

5. если показатели описывают сходные процессы, то их признаковые части совпадают,

6. если основание показателя явл вычисляемым по значениям других оснований, то набор признаков такого показателя – это объединение признаков этих оснований.

При хранении экономических показателей в памяти ЭВМ 1 файл выделяется под группу показателей с одинаковыми реквизитами-признаками.

Формула показателя

П: {(О, С. П, Ф), (У, Е, В), Осн}

О – объект,

С - субъект,

П – процесс, происходящий над объектом,

Ф – формализованная характеристика действия,

У – функция управления,

Е – единицы информации,

В – временные характеристики,

Осн – основание показателя (количественного типа).

Показатель может иметь полный состав признаков, а также какую-то их часть, но основание присутствует всегда.

Виды показателей

1. статический (констатация уже произошедшего факта),

2. динамический (в зависимости от временного фактора показатель меняется).

По способу вычисления:

1. абсолютные (получаемые прямым счётом),

2. относительные (получаемые на основе соотношения других показателей),

С точки зрения автоматизации обработки:

1. переменные (являются исходными для последующих данных),

2. условно-постоянные (данные, которыми пользуются в течение какого-либо времени и которые не меняются) – например. Каждому предприятию присваивается № для налоговой инспекции.

Пример

Атрибуты документа «Приходный ордер»

1 Дата	3 К мат – код материала	5 Цена
2 Склад	4 К-во док – кол-во по документу	6 Сумма
7 Пост – код поставщика	8 К-во пр – кол-во принято	

Атрибутами-основаниями явл:

4 К-во док,

8 К-во пр,

5 Цена,

6 Сумма, которые представляют количественную характеристику процесса оприходования материала на складе.

Таким образом, в документе 4 показателя – по 1 на каждое основание.

Выяснение структуры каждого показателя связано с определением атрибутов-признаков для соответствующих оснований.

У основания **4 К-во док** необходимыми признаками будут **3 К мат, 2 Склад, 7 Пост** (склад принимает материал от конкретного поставщика), **1 Дата.**

В результате структура показателя П 1 принимает вид

П 1 (4 3271).

При рассмотрении П 2 с основанием **8 К-во пр можно использовать правило** (описание сходных процессов).

П 2 (8 3271).

Для показателя П 3 с основанием **5 Цена,** необходимо установить зависят ли цены материалов от предприятия-поставщика или они постоянны. Если допустить последнее, то получим **П 3 (5 3).**

6 Сумма в показателе П 4 является результатом вычисления:

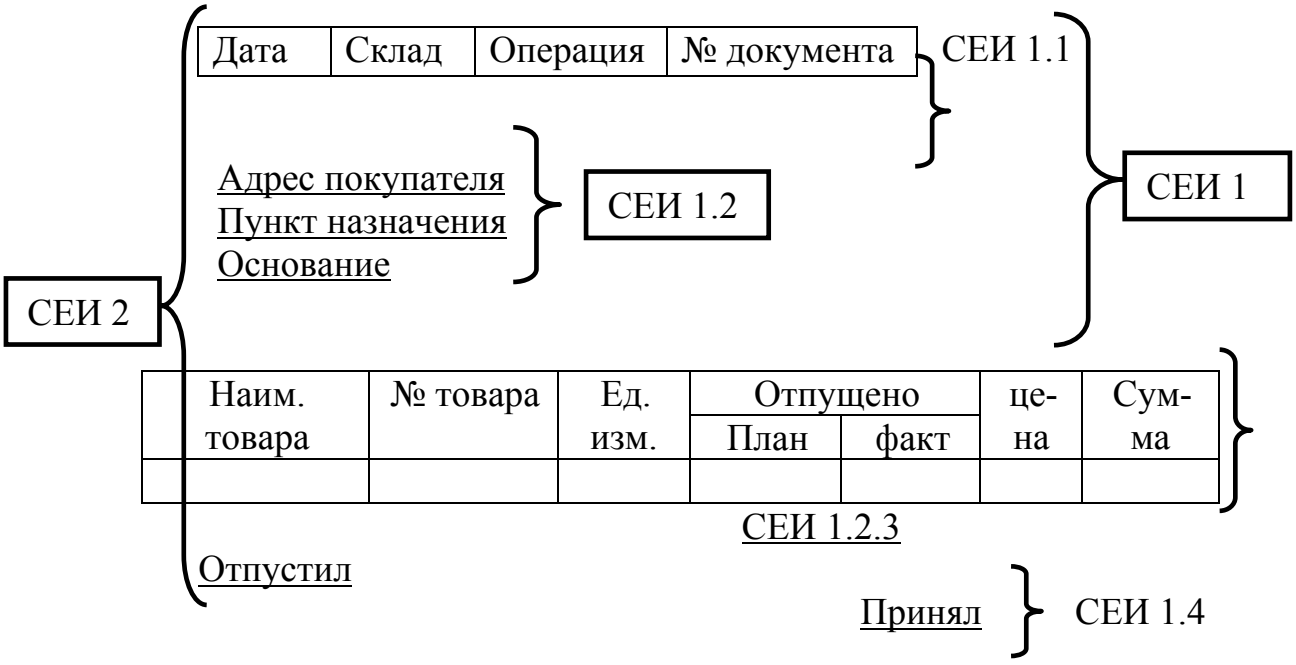
Сумма = К-во пр* Цена, поэтому (согласно правилу)

П 4 получаем на основе П 2 и П 3, то есть

П 4 (6 3271).

Рассмотрим документ (СЕИ)

Неизменная часть документа «Накладная на отпуск продукции»



Стандартные части документа

1. Общая – зона однозначных реквизитов СЕИ 1,
2. Предметная – зона многозначных реквизитов,
3. Оформительская СЕИ 1.4.

Способы формализованного представления СЕИ:

1. табличный (нормализованные СЕИ),
2. графический (представление структуры СЕИ в виде иерархического дерева),
3. аналитический (написание формулы СЕИ).

ТЕМА 2. ЭКОНОМИЧЕСКАЯ ИНФОРМАЦИОННАЯ СИСТЕМА

1. **Структура и основные свойства ЭИС.**
2. **Жизненный цикл ЭИС.**

Внутренняя структура ЭИС

1. БД,
2. Метаданные – данные о данных, внутренняя информация,

Метаинформацию следует представлять как информацию об информации.

3. ПО- м/б представлено СУБД или пакетом прикладных программ (ППП),

СУБД – это комплекс программ, обеспечивающих централизованное хранение, накопление, модификацию и выдачу данных. Находящихся в БД.

ППП – специальное ПО. Написанное для конкретной предметной области или для решения определенных задач.

ИС, дополненная прикладными программами различного назначения, образует систему обработки данных. **СОД.**

Предназначена для учета и оперативного регулирования хозяйственных операций, подготовки стандартных документов для внешней среды (счетов, накладных, платежных поручений). Эти задачи имеют итеративный, регулярный характер и выполняются непосредственными исполнителями хозяйственных процессов и связаны с оформлением и пересылкой документов в соответствии с четко определенными алгоритмами. Результаты выполнения хозяйственных операций через экранные формы вносятся в БД.

ИС, проводящие информационное обслуживание специалистов, принимающих решения, становятся автоматизированной системой управления (**АСУ**). Должна осуществлять выбор управленческих решений (автономно или с участием специалиста). Прикладные программы АСУ, формирующие управленческие решения, как правило, используют экономико-математические методы для выбора оптимальных решений.

Типичными для АСУ являются задачи оптимального управления запасами материалов и полуфабрикатов на складах предприятия, анализ и планирование поставок. Задачи решаются на основе накопленной базы оперативных данных.

4. Вычислительная система – одна или комплекс ЭВМ, связанные м/д собой каналами связи.

Предметная область – это элементы материальной системы , информация о которых хранится и обрабатывается в ЭИС.

Информационным отображением всей предметной области экономического объекта служит информационная база ЭИС.

Для описания Предметной области необходимы 4 понятия:

1. объект,
2. свойство объекта,
3. взаимодействие объектов,
4. свойство взаимодействия.

В ЭИС понятие **объекта** сужается до понятия физического объекта, под которым понимают любой предмет, занимающий место в пространстве.

Отдельный предмет является **экземпляром** объекта, множество предметов – **тип** объекта.

Объекты экономической сферы группируются в 3 крупных типа:

- средства производства,
- предметы труда.
- исполнители.

Свойство объекта – некоторая величина, характеризующая объект в любой момент времени (отдельный экземпляр объекта можно точно описать, если указать достаточное количество значений его свойств. Два экземпляра являются различными, если отличаются по значению хотя бы одного свойства).

Взаимодействие объектов – участие нескольких объектов в каком-либо процессе, кот протекает во времени и пространстве.

Свойство взаимодействия – такое свойство, которое характеризует совместное поведение объектов, но не относится ни к одному из объектов в отдельности.

Например, при производстве изделий взаимодействуют объекты: рабочий, материал, оборудование, изделие.

Кол-во изделий, произведенных за определенный день явл свойством взаимодействия, но никак не характеризует указанные выше объекты, взятые в отдельности.

Среди свойств, описывающих объект выделяют

Идентифицирующие св-ва, то есть св-ва, по значению которых можно однозначно отличить данный экземпляр объекта от какого-либо другого.

Функции ЭИС

1. передача информации,
2. регистрация (первичный учет) информации,
3. отбор необходимой информации,
4. хранение информации,
5. ввод данных,
6. обработка данных,
7. вывод результатной информации

Три составные части обработки информации (этапы)

I наблюдение за объектом управления; регистрация первичных данных, получаемых из различных документов; выбор модели обработки информации;

II сбор и обработка данных с целью получения сводной обобщенной информации в форме, удобной для пользователя;

III получение результатной информации, её анализ; изучение и формирование на её основе управленческих воздействий на объект управления.

Требования к ЭИС

1. Полнота информации.
2. Достоверность.
3. Достаточность.
4. Своевременность.
5. Экономичность (экономические затраты на обработку информации в ЭИС д/б меньше экономич выигрыша на объекте при использовании этой информации).
6. Адаптивность (способность ЭИС изменять свою структуру и закон поведения для достижения оптимального результата при изменяющихся внешних условиях).

Стадии ЖЦ ЭИС

1. Стадия проектирования (детальное описание всех компонентов ЭИС для её последующей реализации) - процесс составления описания еще несуществующей системы на разных языках и с различной степенью детализации, в ходе которого осуществляется оптимизация проектных решений.

ТЗ (постановка задачи), ТЭО (необходимость решения задачи),

Выбор модели данных, технических средств. ПО,

Рабочее проектирование – на основе выбранных и принятых решений реализуются функции системы. Строятся технологические процессы решения задачи; разрабатывается руководство (инструкция) пользователя; решение тестового примера

2. Внедрение (эксплуатация ИС). Период стабильного функционирования ЭИС, не требующий изменения ранее принятых проектных решений.

Если не возникает никаких конфликтов, и все функции реализуются, то система поставляется в промышленную эксплуатацию.

3. модернизация и утилизация системы. Процесс корректировки проектных решений по отдельным компонентам системы.

Существующая система дополняется, но если изменений > 50%, то создается новая. Средний ЖЦ ЭИС 5 лет

ТЕМА 3. МОДЕЛИ ДАННЫХ РЕЛЯЦИОННАЯ МОДЕЛЬ ДАННЫХ

1. **Основные понятия.**
2. **Функциональные зависимости и ключи отношений.**
3. **Операции над отношениями.**
4. **Нормализация отношений.**

Модель – это взаимозависимые сведения о единицах информации, то есть об их структуре, имени и множестве значений.

Модель данных – это формализованный аппарат, который устанавливает:

- **допустимые информационные конструкции,**
- **допустимые операции,**
- **допустимые ограничения, которым подвергается система.**

Допустимые операции:

- проекция,
- выборка,
- соединение,
- деление,
- объединение,
- пересечение,
- вычитание.

Ограничения:

- функциональные зависимости между атрибутами отношения

Основными понятиями РМД являются:

1. тип данных,
2. домен,
3. атрибут,
4. кортеж,
5. ключ.

Понятие тип данных полностью эквивалентно соответствующему понятию в алгоритмических языках.

Домен – это множество значений атрибута (реквизита) – столбец.

Декартово произведение доменов представляет собой множество кортежей.

Кортеж – элементы отношения (строка таблицы).

Отношение – это подмножество декартового произведения списка доменов.

Эта модель является наиболее абстрактной, т.к. она в большей степени ориентирована на конечного пользователя.

Объекты и атрибуты находятся м/д собой в определенных отношениях:

- логические,
- арифметические (бинарные однородные отношения м/д 2-мя отдельными значениями, совпадающими по структуре элементов),

- множественные (отношения м/д однородными множествами элементов; можно применять операции соединения, выборки, и т.д.);
- отношения м/д неоднородными множествами элементов:

= отношение1:1 – однозначное соответствие элементов; происходит тогда, когда одному элементу предметной области ставится в соответствие только 1 другой элемент этой предметной области;

= отношение1:m – одному элементу ставится в соответствие множество элементов другой предметной области;

= m:m – многозначное соответствие; одному элементу одного множества ставится в соответствие множество элементов другого множества и наоборот.

- Семантические отношения – означают смысловую роль элементов в конкретном сообщении (информации) и могут быть использованы для формализации, передачи смысла этой информации.

Отношения в РМД представлены в виде таблицы.

Свойства отношений

1. отсутствие повторяющихся строк,
2. порядок строк несущественен,
3. -//-//- столбцов -//-//-,
4. все значения имеют атомарный (единичный) характер, их нельзя разбить на компоненты без потери информации.

К основным **достоинствам реляционного подхода** следует отнести:

1. наличие небольшого набора абстракций, которые позволяют сравнительно просто моделировать большую часть распространенных предметных областей и допускают точные формальные определения, оставаясь интуитивно понятными;
2. наличие простого и в то же время мощного математического аппарата, опирающегося главным образом на теорию множеств и математическую логику и обеспечивающего теоретический базис реляционного подхода к организации БД;
3. возможность манипулирования данными без необходимости знания конкретной физической организации БД во внешней памяти.

Недостатки:

1. некоторая ограниченность при использовании в так называемых нетрадиционных областях (системы автоматизации проектирования),в которых требуются предельно сложные структуры данных; что явл прямым следствием простоты.
2. невозможность адекватного отражения семантики предметной области – возможности представления знаний о семантической специфике предметной области в реляционных системах очень ограничен.

Таблица

Пример отношения СТУДЕНТЫ реляционной БД

№_студенческого_ билета	Имя	Дата_ рождения	Курс	Специальность
23980282	Алексеев Д.А.	12.03.1982	2	Биология
22991380	Яковлев Н.В.	25.12.1979	4	Физика
22657879	Михайлов В.В.	29.02.1979	5	Математика
24356783	Афанасьев А.В.	19.08.1983	1	Иностранный язык
24350283	Кузнецов В.И.	03.10.1982	1	Физика
23125681	Смирнов А.Д.	26.03.1981	3	История

В рассматриваемом примере используются 3 типа данных – строковый (столбцы «Имя» и «Специальность»), временной тип (столбец «Дата_рождения») и целочисленный тип («Курс», «№_студенческого_билета»).

- Домены: «Имена», «Специальности» для столбцов «Имя» и «специальность» соответственно будут базироваться на строковом типе данных - в число их значений могут входить только те строки, которые могут изображать имя и название специальности (в частности такие строки не должны начинаться с ь).
- Домен «Даты_рождения» для столбца «Дата_рождения» определяется на базовом временном типе данных – данный домен содержит только допустимый диапазон дат рождения студентов.
- Домены «Номер курса» и «Номера_студенчесиких_билетов» базируются на целочисленном типе – в число его значений могут входить только те целые числа, которые могут обозначать номер курса университета (обычно от 1 до 6) и номер студенческого билета (обязательно положительное число).

Столбцы отношения называются атрибутами, им присваиваются имена, по которым к ним затем производится обращение.

Список имен атрибутов отношения с указанием имен доменов называется **схемой отношения.**

Схема отношения СТУДЕНТ запишется так:

СТУДЕНТ {№_студенческого_билета Номера_студенческих_билетов
Имя Имена,
Дата_рождения Даты рождения,
Курс Номера_курсов,
Специальность Специальности}

Кортеж, соответствующий данной схеме отношения представляет собой множество пар {имя атрибута, значение}, которое включает вхождение каждого имени атрибута, принадлежащего схеме отношения. «Значение» явл допустимым значением домена данного атрибута.

Функциональные зависимости и ключи отношений.

В отношении R (A,B,…) атрибут A функционально определяет атрибут B, если в любой момент времени каждому значению A соответствует единственное значение B (A→B). Иначе говорят, что B функционально зависит от A (B=f(A)).

Каждое из отношений имеет набор ключей.

Вероятный ключ – это минимальный набор реквизитов, которые вместе функционально определяют любой реквизит отношения.

- такое множество атрибутов, что каждое сочетание их значений встречается только в 1 строке отношения, и никакое подмножество атрибутов этим свойством не обладает.

Вероятных ключей может быть несколько.

Систематическая проверка свойств вероятного ключа позволяет следить за достоверностью информации в отношении. Когда в отношении несколько вероятных ключей одновременное слежение за ними очень затруднено и целесообразно выбрать один из них в качестве основного – **первичного.**

Первичный ключ – это такой вероятный ключ, по значениям которого производится контроль достоверности информации. (в отношении м/б только 1 первичный ключ)

Теоремы о функциональных зависимостях:

1. A, B → A, то A, B → B,
2. A → B и A → C, тогда и только тогда, когда A → BC,
3. A → B и B → C, то A → C,
4. Если A → B, то AC → B (C произвольно),
5. Если A → B, то AC → BC,
6. Если A → B и BC → Д, то AC → Д.

Операции над отношениями

Для общения пользователя с РМД используется реляционная алгебра и реляционные исчисления.

Реляционная алгебра – это совокупность бинарных и унарных операций над отношениями.

Реляционные исчисления - это совокупность правил для записи выражения. Определяющего вывод нового отношения.

Основные элементы реляционной алгебры – это отношения, к которым применимы все теоретико-множественные операции (логические операции, сложение, умножение, деление).

Такие бинарные операции выполняются только над совместимыми по объединению отношениями.

Совместимые по объединению – такие отношения, которые имеют одинаковую схему и один и тот же порядок реквизитов. **Порядок** – число атрибутов в каждом отношении.

Операции

1. ОБЪЕДИНЕНИЕ ОТНОШЕНИЙ (F∨P)

Имеется 2 отношения

F склад №1

поставщик	код товара
Фирма 1	1111
Фирма 2	1211
Фирма 3	3456

P склад №2

поставщик	код товара
Фирма 1	1111
Фирма 1	1486
Фирма 4	2244
Фирма 3	3456

Получим отношение L, которое содержит строки, присутствующие либо в отношении R, либо в отношении F.

поставщик	код товара
Фирма 1	1111
Фирма 1	1486
Фирма 2	1211
Фирма 3	3456
Фирма 4	2244

2. ВЫЧИТАНИЕ – это множество кортежей, принадлежащих отношению F и не принадлежащих отношению P.

поставщик	код товара
Фирма 2	1211

3. ПЕРЕСЕЧЕНИЕ – множество кортежей, принадлежащих F и P одновременно.

поставщик	код товара
Фирма 1	1111
Фирма 3	3456

4. ДЕКАРТОВО ПРОИЗВЕДЕНИЕ (F⊕P)

Декартовым произведением отношения F порядка n (F(n)) и отношения P порядка m (P(m)) называется множество кортежей длины n+m, из которых первые n-компонентов ∈F, а последующие m ∈P.

F родители

ФИО
И.И.И.
И.М.И.

P дети

ФИО	Год рожд
И.О.И	1978
И.А.И.	1981

В результате операции получим

ФИО р	ФИО д	Год рожд
И.И.И.	И.О.И	1978
И.И.И.	И.А.И.	1981
И.М.И.	И.О.И	1978
И.М.И.	И.А.И.	1981

5. ПРОЕКЦИЯ (отношения F)

Позволяет изменить число атрибутов и их порядок в каком-то одном отношении. Если необходимо выбрать некоторые домены в отношении, уничтожив при этом другие столбцы, затем удалить из результата повторяющиеся строки, по полученное отношение будет называться **проекцией.**

6. СОЕДИНЕНИЕ

Позволяет соединять в одном отношении кортежи, принадлежащие различным исходным кортежам.

Для определения этой операции вводится понятие **сравнимых атрибутов.**

Атрибут A и B одного и того же или различных отношений называются сравнимыми. Если для каждого значения a (a∈A) и в ∈ B можно записать выражение <a ∅ в> истинно или ложно

<a = в>, <a ≠ в>,

<a > в>, <a>= в>

<a < в>, <a <= в>.

Каждая строка первого исходного отношения сопоставляется по очереди со всеми строками второго отношения, и если для этой пары строк соблюдается условие соединения, то они сцепляются и образуют очередную строку в результирующем отношении.

7. ОГРАНИЧЕНИЕ (F [A∅B])

Если из отношения необходимо выделить те кортежи, которые удовлетворяют заданному условию, то вводится операция ограничения.

В качестве значения атрибута (как правило, В) может быть как конкретное значение, так и какой-либо атрибут.

Нормализация отношений

При использовании РМД существенное значение имеют эксплуатационные характеристики, которые проявляются при внесении изменений в БД.

Для улучшения эксплуатационных характеристик применяется метод преобразования отношений – **нормализация.**

Центральная задача проектирования БД – определение количества отношений (или иных СЕИ) и их атрибутного состава. Рациональные варианты группировки реквизитов в отношениях должны учитывать следующие требования:

- Множество отношений должно обеспечивать минимальную избыточность представления информации,
- Корректировка отношений не должна приводить к двусмысленности или к потере информации,
- Перестройка набора отношений при добавлении в БД новых атрибутов должна быть минимальной.

Нормализацию можно определить как процесс, направленный на уменьшение избыточности информации в РМД.

Ограничения на значения, хранимые в РБД, достаточно многочисленны. Соблюдение этих ограничений в конкретных отношениях связано с наличием так называемых **нормальных форм.**

Процесс преобразования отношений БД к той или иной НФ называется нормализацией.

Нормальными формами называются отношения с допустимыми ограничениями на хранимые данные.

НФ нумеруются от 1 по возрастанию. И чем больше № НФ, тем больше ограничений на хранимые значения должны соблюдаться в соответствующем отношении.

На сегодняшний день известно 6 НФ. Практически применяются 3.

1. **устранение повторяющихся групп** → 1 НФ→ устранение неполных функциональных зависимостей → 2 НФ →устранение транзитивных функциональных зависимостей → 3 НФ → устранение многозначности → 4 НФ (нормализация СЕИ приводит к 1 НФ, а нормализация отношений РБД обычно производится до 3 НФ или 4 НФ).

При переходе к следующей НФ свойства предыдущих НФ сохраняются.

Отношение в 1 НФ – это обычное отношение с двухуровневой структурой (недопустимость в структуре отношения 3-го и последующих уровней является ограничением, определяющим 1 НФ отношения). Отношение представляет собой двухмерную таблицу.

Приведение к 1 НФ не связано с анализом смысла атрибутов, а просто означает приведение отношения к каноническому виду.

Отношение имеет **2 НФ**, если оно находится в 1 НФ и не содержит неполных функциональных зависимостей.

Неполная функциональная зависимость имеет место тогда, когда есть 2 следующие зависимости:

- Многореквизитный ключ многофункционально определяет какой-то неключевой атрибут,
- Один из реквизитов этого многозначного многореквизитного ключа – неключевой атрибут.

(неключевым является атрибут не входящий в состав первичного ключа).

3НФ Правило Если отношение находится не во 2 НФ и не в 3 НФ. то оно разделяется на части с помощью операции «Проекция».

Если в отношении существуют зависимости А→Х, Х→У, то говорят о транзитивной зависимости.

Транзитивные зависимости приводят к аномалиям в работе с отношениями.

Чтобы ликвидировать транзитивную зависимость между неключевыми реквизитами вводится понятие 3 НФ. отношение имеет 3 НФ. если оно находится во 2 НФ и не содержит транзитивной зависимости.

Транзитивная зависимость представляет собой наличие двух видов зависимостей:

1. ключ отношения определяет неключевой атрибут,
2. этот неключевой атрибут определяет другой неключевой атрибут.

студент	группа	факультет
→	→	

Отношения: студент → группа, группа → факультет.

Проведя операцию «Проекция» получим студент → группа, студент → факультет.

Избыточность данных связана с тем, что принадлежность группы к факультету указывается столько раз, сколько студентов обучается в этой группе.

Алгоритм получения отношения в 3 НФ обладает следующими свойствами:

- сохраняет все первоначальные ФЗ, т. е. зависимость, справедливая в N, справедлива и в одном из произвольных отношений.
- обеспечивает соединение без потерь, т. е. значения исходного отношения могут быть восстановлены из проекции исходного отношения (N) с помощью операции соединения.
- Результат декомпозиции в 3 НФ содержит меньше значений атрибутов, чем исходное отношение (происходит уменьшение избыточности).

СЕТЕВАЯ МОДЕЛЬ ДАННЫХ (СМД)

1. **Основные понятия СМД.**
2. **Операции с сетевой БД.**

СМД базируется на сетевых структурах, которые возникли из-за необходимости в процессе формирования выходных документов обрабатывать сразу несколько информационных массивов. Это привело к установлению перекрестных ссылок между массивами.

СБД представляет собой множество отношений и веерных отношений. Отношения разделяются на основные и зависимые.

Веерное отношение состоит из 1 основного и 1 зависимого отношения и связей между ними, при условии, что каждое значение зависимого отношения связано с единственным значением основного отношения.

Если существует веерное отношение, то ключ зависимого отношения функционально определяет ключ основного отношения, и наоборот: функциональная зависимость ключей определяет наличие веерного отношения.

Ключом отношения называют атрибут или группу атрибутов, которые функционально определяют каждый из атрибутов отношения.

СБД в зависимости от накладываемых на них ограничений разделяются на:

- Двухуровневые сети,
- Многоуровневые сети.

Ограничение двухуровневых сетей состоит в том, что каждое отношение может существовать в одной из перечисленных ролей:

- вне каких-либо веерных отношений,
- в качестве основного отношения в любом количестве веерных отношений,
- в качестве зависимого отношения в любом количестве веерных отношений.

Также вводятся дополнительные ограничения:

- отношение, которое является основным в одном веерном отношении не м/б зависимым в другом веерном отношении,
- ключ основного отношения м/б только одноатрибутным,
- веерное отношение существует, если ключ основного отношения является частью ключа зависимого отношения.

Многоуровневые сети не предусматривают никаких ограничений на взаимосвязь веерных отношений.

СМД м/б представлена в двух видах:

1. табличный (основными категориями являются записи и связи),
2. графический (модель представляется графом, вершинами которого являются данные об объектах и их атрибутах. Дуги графа – это связи м/д объектами и атрибутами).

Организация веерного отношения в памяти ЭВМ.

В структуру основного и зависимого отношения вводится дополнительный атрибут, называемый *адресом связи – атрибут в составе записи, в котором хранится начальный адрес или номер следующей обрабатываемой записи*.

Значения адресов связи обеспечивают в веерном отношении соответствие каждого значения зависимого отношения единственному значению основного отношения.

Связь значений зависимого отношения с единственным значением основного отношения обеспечивается следующим образом:

Адрес связи некоторой записи основного отношения указывает на одну из записей зависимого отношения, адрес связи указанной записи зависимого отношения – на следующую запись зависимого отношения, связанную с той же записью основного отношения и т.д.

Последняя запись зависимого отношения в этой цепочке адресует эту запись на запись основного отношения.

Полученная кольцевая структура связи называется **веером**.

Преимущества СМД

1. универсальность.
2. возможность доступа к данным через значения нескольких отношений (например, через любые основные отношения),

Недостатки

1. сложность, т.е. обилие понятий, вариантов их взаимосвязей и особенностей реализации,
2. допустимость только навигационного принципа доступа к данным.

(центральным является понятие «текущая запись» в отношения БД. Текущей записью в отношениях после выполнения некоторой операции является значение отношения, на котором операция завершилась. Следующая операция начинается с этой текущей записи, а в результате выполнения операции положение текущей записи изменяется (завершение операции может изменить положение текущей записи и в других отношениях)).

Операции с сетевой БД.

1. пересечение,
2. объединение,
3. вычитание, (рассмотрены выше в РМД)
4. образ вычисляет образ каждого значения атрибута или показателя и производит объединение полученных образов.

В отношении T (A,B) образом значения a атрибута A является множество значений атрибута B, и каждый элемент b этого множества образует вместе с a некоторую строку (или часть строки) отношения T.

im B(a) = {b1, b2,…bk},

где im – обозначение операции «образ»,

a – значение, образ которого вычисляется.

B – имя атрибута для образа значения a,

b1, b2,…bk – значения атрибута B.

Существует отношение U (ФИО, ЯП), где для каждого программиста указываются языки программирования, которые он знает.

U	
ФИО	ЯП
Иванов	Си
Иванов	Фортран
Иванов	Паскаль
Петров	Си
Петров	Паскаль
Сидоров	Си
Сидоров	Фортран
Николаев	Фортран
Николаев	Паскаль

im ФИО («Си») = {«Иванов», «Петров», «Сидоров»}

im ФИО («Фортран») = {«Иванов», «Сидоров», «Николаев»}

im ФИО («Си») = {«Иванов», «Петров», «Сидоров»}

5. пересечение образов

вычисляет образы значений данных и производит пересечение этих образов.

im ФИО («Си») ∩ im ФИО («Фортран»)= {«Иванов», «Сидоров»}

6. сечение

позволяет извлечь из показателей (записи) один из компонентов при условии того, что показатель образован сочетанием нескольких.

7. нахождение max и min по значениям ключа.

ИЕРАРХИЧЕСКАЯ МОДЕЛЬ ДАННЫХ

1. **Основные понятия ИМД.**
2. **Ограничения.**
3. **Операции.**

ИМД имеет много общего с СМД (хронологически она появилась даже раньше). Иерархический подход обеспечивает естественный способ моделирования предметной области. (*Его эффективно применять, если структура предметной области соответствует условиям задачи классификации.)* Допускается отображение одной предметной области в нескольких ИБД.

Допустимыми информационными конструкциями являются:

- отношение,
- веерное отношение,
- ИБД.

ИБД называется множество отношений и веерных отношений, для которых соблюдаются 2 ограничения:

1. существует единственное отношение, называемое корневым, которое не является зависимым ни в одном веерном отношении.
2. все остальные отношения являются зависимыми отношениями только в одном веерном отношении.

ИМД представляет собой графовую модель с вершинами – таблицами. Структурная диаграмма ИМД представляет собой *упорядоченное иерархическое дерево*, в котором определено относительное расположение вершины и дуг соответствующим функциям в виде связи, направленной от корней к листьям.

Вершины дерева – это СЕИ, которые в ИМД называются **сегментами.**

Дуги – это связи исходного и порожденного сегментов.

Иерархический путь – это последовательность сегментов; начинается от корневого сегмента, в котором последующие сегменты выступают попеременно в ролях исходного и порожденного.

Запись иерархической позиции – это совокупность одного значения корневого сегмента вместе со всеми значениями других сегментов, присутствующих в иерархическом пути.

Правило. Число различных записей в ИМД = числу различных значений корневого сегмента.

Совокупность записей ИМД, порожденных одним корневым сегментом, образуют 1 ИБД.

В ИМД реализуется связь 1:m. (например, преподаватель: дисциплина).

Ограничения ИМД.

1. типы связей д/б функциональными,
2. структура связей д/б древовидной,

Операции:

1. Получение уникальной записи.

Позволяет выделить первое из значений некоторого сегмента, удовлетворяющее сформированным условиям. Каждое условие относится к одному из сегментов, лежащих на иерархическом пути между корневым и искомым сегментом.

Правило. ***Если в веерном отношении ИМД один и тот же атрибут присутствует и в основном и в зависимом отношении, то из зависимого отношения такой атрибут следует исключить.***

2. Получение следующей записи.

Этот оператор предназначен для выборки следующей записи после той, на которой остановилось действие оператора.

3. Получение следующей внутри записи.

Получение следующего значения внутри записи.

Достоинства ИМД

1. Простота (хотя модель использует 3 информационные конструкции, иерархический принцип соподчиненности понятий является естественным для многих экономических задач).
2. Минимальный расход памяти.

Недостатки ИМД

1. Неуниверсальность
2. Допустимость только навигационного доступа к данным.
3. Доступ к данным производится только через корневое отношение.

ТЕМА 4. МЕТОДЫ ОРГАНИЗАЦИИ ДАННЫХ ОРГАНИЗАЦИЯ ДАННЫХ И АНАЛИЗ АЛГОРИТМОВ

1. **Основные положения.**
2. **Последовательная организация данных.**
3. **Цепная организация данных.**
4. **Древовидная организация данных.**

Методы организации данных (МОД) в памяти ЭВМ обычно предполагают раздельное хранение значений каждой СЕИ. Отдельное значение СЕИ, находящееся в памяти ЭВМ, называется ***записью.*** Запись состоит из значений атрибу-

тов, входящих в структуру СЕИ. Множество записей образует **массив** или **файл.** Термин массив обычно используют при рассмотрении данных в памяти ЭВМ, а термин файл применяется для данных, хранимых на внешних запоминающих устройствах. (ВЗУ) Как правило. Файл содержит записи, принадлежащие одной и той же СЕИ, хотя в общем случае это не является обязательным.

Под организацией значений данных понимают относительно устойчивый порядок расположения записей данных в памяти ЭВМ и способ обеспечения взаимосвязи м/д записями.

Организация значений данных м/б **линейной и нелинейной**. При линейной организации данных каждая запись. Кроме первой и последней, связана с одной предыдущей и одной последующей записями. У записей, соответствующих нелинейной организации данных, количество предыдущих и последующих записей может быть произвольным.

Линейные методы организации данных различаются только способами указаний предыдущей и последующей записи. Но это приводит к тому, что алгоритмы. Эффективные для одних методов организации данных, становятся неприемлемыми для других методов.

Среди линейных методов выделяются **последовательная и цепная** организация данных.

При *последовательной организации данных* записи располагаются в памяти строго одна за другой, без промежутков, в той последовательности, в которой они обрабатываются. Последовательная организация данных обычно и соответствует понятию массив (файл).

Записи, составляющие массив, с точки зрения способа указания их длины делятся на записи фиксированной, переменной и неопределенной длины. **Записи фиксированной (постоянной**) длины имеют одинаковую, заранее известную длину. Если длины записей неодинаковы, то длина указывается в самой записи. Такие записи называются **записями переменной длины.**

Вместо явного указания длины записи можно отмечать окончание записи специальным символом-разделителем, который не должен встречаться среди информационных символов значения записи. Записи, заканчивающиеся разделителем, называются **записями неопределенной длины.**

Адреса промежуточных записей фиксированной длины в массиве задаются формулой

A (i) = A (1)+(I - 1)*L,

где A (1) – начальный адрес первой записи,

A (i) – начальный адрес i-й записи,

L – длина одной записи.

(Для массива переменной длины такой формулы просто не существует. Они занимают меньший объем памяти, но их обработка ведется с меньшей скоростью, поскольку затруднено обнаружение следующей записи.)

В структуре записей последовательного массива обычно выделяется один или несколько ключевых атрибутов, по значениям которых происходит доступ к остальным значениям атрибутов той или иной записи. Состав ключевых атри-

бутов необязательно соответствует понятию первичного ключа.

Ключевые атрибуты в записях обозначаются через p(i), где I – номер записи, общее число записей в массиве обозначается через M.

Записи массива могут быть упорядоченными или неупорядоченными по значениям ключевого атрибута (ключа), имя которого одинаково во всех записях. Ключевой атрибут обычно является *атрибутом-признаком*. Часто требуется поддерживать упорядоченность записей по нескольким именам ключевых признаков. В этом случае среди признаков устанавливается старшинство. Условие упорядоченности записей в массиве (и вообще для линейной организации данных) выглядит следующим образом:

p(i) < = p(i+1) – упорядоченность по возрастанию,

p(i) > = p(i+1) – упорядоченность по убыванию.

Наиболее важными и часто применяемыми алгоритмами обработки данных являются формирование данных, их поиск и корректировка, а также последовательная обработка. Эти алгоритмы могут быть реализованы с использованием достаточно большого количества методов организации данных. Здесь мы рассмотрим выбор наилучшего метода организации данных для названных алгоритмов. Сами методы организации данных будут представлены в их простейшей форме.

Данные обычно возникают в неупорядоченной форме. Перед обработкой, как правило, целесообразно упорядочить их по значениям ключевых атрибутов, что составляет одну из основных работ по формированию (подготовке) данных. Процедуру упорядочения файла часто называют **сортировкой**.

Упорядоченные данные эффективны для организации быстрого поиска информации. Выходные документы, выводимые на печать, полученные на основе отсортированных данных, удобны для дальнейшего использования. Многие алгоритмы задач управления вообще рассчитаны на использование только упорядоченных данных. Отсортированные данные позволяют организовать быструю обработку нескольких массивов. (Далее будем считать все массивы упорядоченными по возрастанию значений одного атрибута, когда для ключа i–й записи p(i) справедливо условие p(i) < = p(i+1).)

Критерии эффективности алгоритмов

Естественной характеристикой эффективности того или иного алгоритма служит время его выполнения в зависимости от ряда параметров хранимой информации. Поэтому для каждого метода организации данных требуется анализировать следующие величины:

- время формирования данных, т.е. время создания данных в памяти ЭВМ так или иначе упорядоченного представления данных (упорядочение способно ускорить выполнение алгоритмов поиска данных);
- время поиска данных. Как известно, условия поиска (выборки) на практике могут быть достаточно разнообразные. Анализируется обычно простейший и наиболее распространенный случай (поиск по совпадению) – найти записи, у

- время корректировки данных. Из всех возможных вариантов корректировки учитывается включение и исключение одной записи;
- объем дополнительной памяти, расходуемой под служебную информацию (например, адреса связи).

При анализе алгоритмов необходим еще ряд допущений, обеспечивающих использование равномерного распределения вероятностей для всех случайных величин, описывающих работу алгоритма, в том числе:

- распределение значений ключевых атрибутов в массиве из M записей – равномерное,
- значение q при поиске по совпадению выбрано случайно: это означает, что поиск с одинаковой вероятностью 1/M может закончиться на любой записи массива,
- положение включаемой (выключаемой) записи при корректировке определяется теми же вероятностями, что и при поиске.

Цепная (списковая) организация данных

Решение целого ряда задач обработки данных требует применения таких методов организации данных, которые позволили бы связать физически разнесенные в памяти данные в логическую последовательность, определяющую порядок их обработки. Простейшим методом, применяемым для этих целей, является **списковая (цепная) организация данных**.

Списком называется множество записей, занимающих произвольные участки памяти, последовательность обработки которых задается с помощью адресов связи.

Адресом связи некоторой записи называется атрибут, в котором хранится начальный адрес или номер записи, обрабатываемой после этой записи.

Обычная последовательность обработки записей в списке определяется возрастанием значений ключа в записях.

Указатель списка – адрес, хранящий положение первой записи.

Конец списка – это специальное значение, отмечающее, что последующей записи нет (0).

Указатель свободной памяти – адрес, хранящий положение первой свободной записи.

Цепной каталог – сплошной участок памяти, в котором одновременно размещаются список обрабатываемых записей и список свободных позиций памяти.

Включение и выключение записей в цепном каталоге предполагает поиск местоположения включаемой (исключаемой) записи и замену значений адресов связи для установления новой последовательности записей основного списка и списка свободной памяти.

Оценка времени корректировки складывается из времени реализации поиска и времени на замену значений адресов связи.

Адрес связи последней записи или последней позиции свободной памяти

отмечается нулевым значением.

Древовидная организация данных.

Деревом называется множество записей, расположенных по уровням следующим образом:

- на 1-м уровне расположена только одна запись (корень дерева),
- к любой записи i-го уровня ведет адрес связи только от одной записи уровня i - 1.

Особенностью этой организации данных является то, что в дереве связаны значения только одного отношения. *(а в иерархической системе связаны различные отношения*).

Количество уровней в дереве называется **рангом**. Ранг определяет максимальное число сравнений при поиске данных. Записи дерева, которые адресуются от общей записи (i - 1)-го уровня, образуют **группу**. Максимальное число элементов в группе называется **порядком дерева**. При размещении дерева в памяти ЭВМ каждая запись может занимать произвольное место. *(Рассмотрим бинарные деревья).*

Бинарные деревья (второго порядка)

Каждый из уровней имеет только 2 выхода**. Особенность –** составляющие его записи могут быть упорядочены. Для этого один из атрибутов записи объявляется ключевым.

Для определения упорядоченности необходимо ввести новые понятия.

Записи, у которых заполнены 2 адреса связи, называются **полными.**

Записи с одним заполненным адресом связи считаются **неполными**. Записи, не имеющие адресов связи, называются **концевыми**.

Каждая из ветвей образует поддерево (левое и правое).

Каждая запись имеет левую и правую ветвь.

ПРАВИЛО

В упорядоченном бинарном дереве значение ключевого атрибута каждой связи должно быть больше, чем значение ключа у любой записи левой её половины (ветви), и меньше, чем ключ любой записи на ее правой ветви.

Упорядоченное бинарное дерево формируется из неупорядоченного массива записей по определенному алгоритму. Этот алгоритм создает дерево из первой записи массива, затем – дерево из первых двух записей, из первых трех записей и так далее до исчерпания всех записей массива.

Алгоритм построения упорядоченного бинарного дерева. (стр. 162)

1. Первая запись массива с ключом p(1) становится корнем дерева.
2. Значение ключа второй записи p(2) сравнивается с p(1). Если p(2) < p(1), то вторая запись помещается на левой от корня ветви. В противном случае – на правой.
3. Выбор места i- ой записи массива производится следующим образом.

Ключ p(i) сравнивается с корневым значением, и выполняется переход по левому адресу (если p(1) > p(i), а при p(1) <= p(i) – по правому адресу. Ключ достигнутой записи также сравнивается с p(i), и снова организуется переход по левому или по правому адресу и т.д. Когда будет достигнут незаполненный адрес связи, то он должен адресовать запись с ключом p(i). Указанные действия повторяются до исчерпания всех записей исходного массива.

ТЕМА 5 ОРГАНИЗАЦИЯ ДАННЫХ ВО ВНЕШНЕЙ ПАМЯТИ ЭВМ

1. **Понятие и классификация файлов.**
2. **Методы организации данных во внешней памяти ЭВМ.**

В качестве внешней памяти ЭВМ используются магнитные диски (для которых характерно примерное равенство затрат времени на чтение и запись).

Время доступа к данным на ВЗУ зависит от места расположения данных на диске или ленте, что существенно отличает их от оперативной памяти и определяет специфику организации данных во внешней памяти ЭВМ.

Данные на ВЗУ хранятся в виде **файлов. Файл представляет собой множество логически связанных записей.**

Файл – это некоторое множество записей однородной структуры, предназначенное для решения экономических задач.

Запись – это набор полей определенного формата, объединенных по общему ключевому полю. Запись обычно соответствует одному значению некоторой СЕИ.

Каждый файл имеет уникальное **имя файла**. В простейшем случае файл представляет последовательный массив записей на ВЗУ.

Все файлы ЭИС можно классифицировать по следующим признакам:

- по этапам обработки (входные, базовые, результатные);
- по типу носителя (на промежуточных носителях – ГМД и лентах и на основных носителях – ЖМД, магнитооптических дисках и т.д.);
- по составу информации (файлы с оперативной информацией и файлы с постоянной информацией);
- по назначению;
- по типу логической организации (файлы с линейной структурой записи, реляционные, табличные);
- по способу физической организации (файлы с последовательным, индексным и прямым способом доступа).

ВХОДНЫЕ файлы создаются с первичных документов для ввода данных или обновления базовых файлов.

Файлы с РЕЗУЛЬТАТНОЙ ИНФОРМАЦИЕЙ предназначаются для вывода ее на печать или передачи по каналам связи и не подлежат длительному хранению.

К числу БАЗОВЫХ файлов, хранящихся в информационной базе (ИБ), относятся:

1. основные;
2. рабочие;
3. промежуточные;
4. служебные;
5. архивные.

1. Основные файлы должны иметь однородную структуру записей и могут содержать записи с оперативной и условно-постоянной информацией. Оперативные файлы могут создаваться на базе одного или нескольких входных файлов и отражать информацию одного или нескольких первичных документов. Файлы с условно-постоянной информацией могут содержать справочную. Расценочную, табличную и другие виды информации, изменяющейся в течение года не более чем на 40%. Файлы со справочной информацией должны отражать все характеристики элементов материального производства (материалы, сырье, основные фонды, трудовые ресурсы и т.д.). Как правило, справочники содержат информацию классификаторов. Нормативно-расценочные файлы должны содержать данные о нормах расхода и расценках на выполнение операций и услуг. Табличные файлы содержат сведения об экономических показателях, считающихся постоянными в течение длительного времени. Плановые файлы содержат плановые показатели, хранящиеся весь плановый период.

2. Рабочие файлы создаются для решения конкретных задач на базе основных файлов путем выборки части информации из нескольких основных файлов с целью сокращения времени обработки.

3. Промежуточные файлы отличаются от рабочих тем, что они образуются с целью дальнейшего использования для решения других задач. Эти файлы, как и рабочие, при высокой частоте обращений могут быть также переведены в категорию основных файлов.

4. Служебные файлы предназначаются для ускорения поиска информации в основных файлах и включают в себя справочники, индексные файлы и каталоги.

5. Архивные файлы содержат ретроспективные данные из основных файлов, которые используются для решения аналитических. Например, прогнозных задач. Архивные данные могут также использоваться для восстановления ИБ при разрушениях.

Методы организации данных во внешней памяти ЭВМ.

Анализ методов организации данных остается в основном справедливыми ля данных во внешней памяти ЭВМ, однако серьезным фактором, влияющим на время доступа, становится взаимное расположение файлов и записей на магнитном носителе.

Определим адресное расстояние dA как разность адресов предыдущего и текущего обращения к запоминающему устройству, взятую со знаком +.

$$dA = |A(i - 1) – A(i)|$$

Чтобы применять адресное расстояние ко всем типам запоминающих устройств, нужно учесть, что с магнитного диска читается (записывается) не отдельный символ (байт), а сектор или блок данных размером, например, 512 байт.

Организация внешней памяти персональных ЭВМ имеет ряд отличий от принципов, используемых в мини-ЭВМ и средних ЭВМ. Вся внешняя память разделена на физические блоки (секторы), имеющие фиксированный размер (обычно 512 байт), который не зависит от желания проектировщика системы. Обмен с оперативной памятью происходит только целыми секторами.

Когда производится только последовательная обработка файла, оптимальный (с точки зрения минимального времени доступа) размер блока должен быть наиболее крупным из возможных; когда происходит только выборка одиночных записей, оптимальными являются блоки размером в одну запись.

Существует ряд стандартных методов организации файлов на магнитном диске и соответственно методов доступа к этим файлам.

Среди них:

- последовательная,
- индексно-последовательная,
- индексно-произвольная,
- прямая организация данных.

При *последовательной организации* файла на магнитном диске возможен доступ от только что обработанной записи к последующей записи (по направлению к концу файла). Переход в обратном направлении не возможен, единственный путь состоит в закрытии файла, повторном его открытии и движения к нужной записи в прямом направлении.

Индексно-последовательный файл представляет собой последовательный файл, снабженный индексами.

Индекс – это набор ключей и адресов записей, которые выбираются из основного массива по определенному закону.

На магнитном диске выделяются 3 области:

- первичная,
- индексная,
- область переполнения.

В первичной области помещаются упорядоченные по значениям ключевого атрибута записи, когда файл впервые создается.

В зависимости от размера первичной области могут создаваться 1, 2 или 3 уровня индексов:

- индекс первого уровня отмечает последнюю запись каждой дорожки магнитного диска;
- индекс второго уровня отмечает последнюю запись каждого цилиндра магнитного диска,

Если файл индекса второго уровня достаточно велик по размеру, то для него допускается создание индекса третьего уровня.

Область переполнения предназначена для размещения записей. Включаемых в индексно-последовательный файл. Новые записи связываются в цепочку и размещаются на том цилиндре, при котором ключи новых записей соответствуют интервалу ключей в первичной области этого цилиндра.

Характеристики индексно-последовательного доступа:

1. значения ключей записей должны быть отсортированы;
2. в индекс заносится наибольший ключ для всех записей блока (дорожки);
3. наличие повторяющихся значений ключа недопустимо;
4. эффективность доступа зависит от числа уровней индексации, распределения памяти для размещения индекса, числа записей в файле и размера области переполнения.

Индексно-произвольный доступ получается, если в индекс попадает информация о ключе каждой записи. Записи файла могут быть при этом не упорядочены по значению ключа. Индекс для индексно-произвольного метода доступа практически всегда формируется как многоуровневый. Типичная организация многоуровневого индекса соответствует понятию B-дерева. Нижний уровень B-дерева образуют индексы со ссылкой на каждую запись основного массива. Благодаря использованию адресных ссылок упорядоченность основного массива е обязательна.

Индексы нижнего уровня разделены на страницы, и в конце каждой страницы оставляется резервная память. Последний индекс каждой страницы поступает на страницу предпоследнего уровня B-дерева. Когда эта страница будет почти заполнена индексами, последний из них поступит на страницу более высокого уровня и т.д.

Прямой метод доступа соответствует файлу, который использует адресную функцию вида $i = p - a$

Адресной функцией называется зависимость $i = f(p)$,

где i – номер (адрес) записи,

p – значение ключевого атрибута в записи.

Простейшая адресная функция имеет вид: i = p – a, где a – константа.

Недостаток этой функции – большой объем неиспользуемой памяти.

Для прямого доступа характерны следующие особенности:

- не требуется упорядоченность записей файла;
- наличие повторяющихся значений ключа недопустимо;
- значениям нескольких ключей может соответствовать один и тот же адрес (блок).

При выборе метода организации файла существенное влияние оказывает количество записей, которое должно быть обработано в процессе реализации запроса. Этот параметр называется долей выборки и равен отношению числа требуемых при выборке записей файла к общему числу записей в файле.

Блок данных на внешнем запоминающем устройстве обычно не заполняется полностью, т.е. оставляется резервная память (обычно 10-15% размера блока). Если этого не делать. То включение новых записей потребует создания для них новых блоков практически при каждой корректировке. Эти блоки будут содержать довольно мало записей, от чего резко возрастет объем дополнительной памяти, необходимый для массива.

Когда резервная память блока будет исчерпана и в него потребуется включить новую запись, наступает переполнение блока.

Частота переполнения описывается формулой:

$$K = (V + 1)/(2p – 1);$$

где K – ожидаемое число корректирующих обращений (включений и исключений записей) к одному блоку до наступления переполнения.

V – объем свободной памяти блока, выраженный в количестве записей;

p >0.5 – вероятность того, что корректирующее обращения является включением.

Если p<= 0,5, то блок, как правило, никогда не переполнится. После переполнения блока вслед за ним в память включается новый блок, в который переписывается половина записей из переполненного блока.

Список литературы

1. Информатика: Учебник. – 3-е перераб. Изд. / Под ред. Н.В. Макаровой. – М.: финансы и статистика, 2002.
2. Информационные системы/Петров В.Н. - СПб:Питер, 2003. (учебник)
3. Информационный менеджмент: Уч. пособие для ВУЗов – М: - ЮНИТИ – ДАНА
4. Мишенин А.И. "Теория экономических информационных систем" Москва 1999г. учебное пособие
5. Смирнова Г.Н. Проектирование экономических информационных систем. Москва, 2001

Содержание