

Информационные системы

Оглавление

1. Информационные системы ([1], п.12.1)
 2. Классификация ИС ([1], п.12.1)
 3. Структура ИС ([1])
 4. Основы применения инструментальных средств ИТ ([1], п.12.4)
 5. Создание автоматизированных информационных систем ([1], п.12.5)
 6. Применение информационных систем
 - 6.1. В науке (обзорно)
 - 6.1.1. Автоматизированные системы научных исследований (АСНИ) (По классификации 2.1.).
 - 6.1.2. Системы автоматизированного проектирования (САПР) (По классификации 2.2.).
 - 6.1.3. Какая взаимосвязь между АСНИ и САПР?
 - 6.2. В обучении (обзорно)
 - 6.2.1. Автоматизированные обучающие системы (АОС) (По классификации 2.5.).
 - 6.3. ИС управления ([1], п.12.2, 12.6) (обзорно)
 - 6.3.1. Основные концепции построения ИС управления (ИСУ) ([1], п.12.2).
 7. Информационно-поисковые системы (По классификации 3.1.).
 - 7.1. Состав Информационно-поисковые системы
 - 7.2. БАНКИ (БАЗЫ) ИНФОРМАЦИИ
 - 7.3. ИНДЕКСАЦИЯ ДАННЫХ
 - 7.3.1. КЛАССИФИКАТОРЫ
 - 7.3.2. Информационно-поисковые языки (ИПЯ) (+ [1], п.7.1)
 - 7.4. Система индексирования ([1], п.7.4)
 - 7.5. Информационный поиск
 8. Поиск в Интернете
 - 8.1. Поисковая система
 - 8.1.2. Поисковая машина
 - 8.1.3. Поисковый робот
 - 8.2. История
 - 8.3. Популярные поисковые системы
 - 8.4. Язык поисковых запросов
 - 8.5. Обзор поисковых систем
 - 8.6. Вопросно-ответная система
 - 8.7. Коллаборативная, или Совместная фильтрация
 - 8.8. Электронная библиотека
 - 8.9. Проблемы в работе поисковых систем
 - 8.9.1. Глубокая паутина
- Информационные источники

1. Информационные системы ([1], п.12.1)

Все системы можно разделить на две основные категории:

1. Материальные системы - это совокупность материальных объектов.

Среди них выделяют:

1.1.технические,

1.2.эргатические (эргатическая система (от греч. ergates - действующее лицо)— сложная система управления, составным элементом которой является человек-оператор (или группа операторов), напр. система управления самолетом, диспетчерская служба вокзала, аэропорта.

1.3.смешанные системы.

Среди смешанных систем особого внимания заслуживает подкласс эргатехнических систем (систем «человек машина»), состоящих из человека-оператора эргатической составляющей и машины (машин) - технической составляющей.

2. Абстрактные системы представляют собой продукт человеческого мышления - знания, теории, гипотезы.

Информационные системы следует относить к категории *материальных*, учитывая при этом, что продукт труда в данных системах нематериален.

Информационная система (ИС) — совокупность взаимосвязанных компонентов, которые обеспечивают средства для протекания **информационных процессов** (получение, хранение, передача, обработка информации).

Одним из назначений ИС является снабжение работников различного ранга информацией для реализации функций управления.

В информационную систему данные поступают от источника информации. Эти данные отправляются на хранение либо претерпевают в системе некоторую обработку и затем передаются потребителю.

ИС создается для конкретного потребителя. Эффективная ИС принимает во внимание различия между уровнями управления, сферами действия, а также внешними обстоятельствами и дает каждому уровню управления только ту информацию, которая ему необходима для эффективной реализации функций управления.

Внедрение ИС производится с целью повышения эффективности производственно-хозяйственной деятельности фирмы не только за счет обработки и хранения рутинной информации, автоматизации конторских работ, но и за счет принципиально новых методов управления, основанных на моделировании действий специалистов фирмы при принятии решений (методы искусственного интеллекта, экспертные системы и т.п), использование современных средств телекоммуникаций (электронная почта, телеконференция), глобальных и локальных вычислительных систем и т.д..

Между потребителем и собственно информационной системой может быть установлена обратная связь. В этом случае информационная система называется *замкнутой*. Канал обратной связи необходим, когда нужно учесть реакцию потребителя на полученную информацию.

2. Классификация ИС ([1], п.12.1)

Сама идея информационных систем и некоторые принципы их организации возникли задолго до появления ЭВМ. Библиотеки, архивы, адресные бюро, телефонные справочники, словари - все это информационные системы.

1. По степени автоматизации (механизации)

1.1. **Ручные ИС** характеризуются тем, что операции по переработки выполняются человеком без применения каких-либо технических средств.

Например, о деятельности менеджера в фирме, где отсутствуют компьютеры, можно говорить, что он работает с ручной ИС.

1.2. **Механизированные ИС** - для выполнения некоторых процедур используются технические средства.

1.3. **! Автоматизированные информационные системы (АИС)** – некоторые функции (подсистем) управления и обработки данных осуществляется автоматически.

Главная роль отводится компьютеру. Это наиболее популярный класс ИС.

1.4. **Автоматические ИС** - все функции управления и обработки данных осуществляется техническими средствами автоматически без участия человека. (человек как звено управления отсутствует). Человек

может выполнять лишь функции внешнего наблюдения за работой системы.

Например, различные роботы, автоматическое управление технологическими процессами, некоторые поисковые машины Интернет, например Google, где сбор информации о сайтах осуществляется автоматически поисковым роботом и человеческий фактор не влияет на ранжирование результатов поиска.

Обычно термином ИС в наше время называют *автоматизированные информационные системы*. Наиболее эффективными в большинстве сложных систем управления являются автоматизированные информационные системы, включающие в свой состав компьютеры.

Автоматизированные системы обработки данных также называют *автоматизированными системами управления (АСУ)*.

Автоматизированные системы обработки данных, которые имеют специальное программное обеспечение для анализа семантики (смысла) информации и гибкой логической ее структуризации, часто называют *системами обработки знаний (СОЗ)*.

АСУ является эффективной, если соответствует следующим принципам:

- **интеграция** — обрабатываемые данные, однажды введенные в АСОД (автоматизированная система обработки данных). Многократно используются для решения задач. Устраняется дублирование данных и операции их преобразования;
- **системность** - обработка данных в различных разрезах с целью получения информации, необходимой для принятия решений на всех уровнях и во всех функциональных подсистемах управления;
- **комплексность** - механизация и автоматизация процедур преобразования данных на всех стадиях техпроцесса.

ИС можно также классифицировать и по другим признакам:

2. По сфере применения

2.1. **Автоматизированные системы научных исследований (АСНИ)** — программно-аппаратные комплексы, предназначенные для автоматизации деятельности научных работников, научных исследований, испытаний.

Например, анализ статистической информации, управление экспериментом.

2.2. **ИС автоматизированного проектирования (САПР)** — программно-технические системы, предназначенные для автоматизации труда инженеров-проектировщиков и разработчиков новой техники и технологий.

Например, выполнение проектных работ с применением математических методов.

2.3. **ИС организационного управления** — обеспечение автоматизации функций административного (управленческого) персонала.

к этому классу относятся ИС управления как промышленными (предприятия), так и не промышленными объектами (банки, биржи, страховые компании, гостиницы, и т.д.) и отдельными офисами (офисными системами).

2.4. **ИС управления техническими процессами** — обеспечение управления механизмами, технологическими режимами на автоматизированном производстве (например, в металлургии, энергетике и т.п.).

2.5. **Автоматизированные обучающие системы (АОС)** — комплексы программно-технических и учебно-методических средств, обеспечивающих учебную деятельность.

2.6. **Интегрированные ИС** - обеспечение автоматизации большинства функций предприятия.

3. По характеру использования информации

3.1. **Информационно-поисковые системы (ИПС)** — предназначенные для сбора, хранения и выдачи информации по запросу пользователя;

3.2. **Информационно-аналитические системы** — предназначенные для аналитической обработки данных.

3.3. **Информационно-решающие системы** — системы, осуществляющие переработку информации по определенному алгоритму.

- управляющие, результатная информация которых непосредственно участвует в формировании управляющих воздействий
- советующие, предлагающие пользователю определенные рекомендации для принятия решений (системы поддержки принятия решений)

3.4. **Ситуационные центры** (информационно-аналитические комплексы)

4. По архитектуре (степени распределённости) различают:

4.1. **Настольные (desktop), или локальные ИС**, в которых все компоненты (БД, СУБД, клиентские прило-

жения) работают на одном компьютере;

4.2. **Распределённые** (distributed) ИС, в которых компоненты распределены по нескольким компьютерам. Распределённые ИС, в свою очередь, разделяют на:

- файл-серверные ИС (ИС с архитектурой «файл-сервер»); БД находится на сервере (файл-сервере), а СУБД и клиентские приложения находятся на рабочих станциях;
- клиент-серверные ИС (ИС с архитектурой «клиент-сервер»). БД и СУБД находятся на сервере, а на рабочих станциях находятся клиентские приложения.

В свою очередь, клиент-серверные ИС разделяют на: двухзвенные и многозвенные. В **двухзвенных** (two-tier) ИС всего два типа «звеньев»: сервер баз данных, на котором находятся БД и СУБД, и рабочие станции, на которых находятся клиентские приложения. Клиентские приложения обращаются к СУБД напрямую. В **многозвенных** (multi-tier) ИС добавляются промежуточные «звенья»: серверы приложений (application servers). Пользовательские клиентские приложения не обращаются к СУБД напрямую, они взаимодействуют с промежуточными звеньями.

5. По охвату задач (масштабности)

- 5.1. **Персональная** информационная система предназначена для решения некоторого круга задач одного человека.
- 5.2. **Групповая** информационная система ориентирована на коллективное использование информации членами рабочей группы или подразделения.
- 5.3. **Корпоративная** информационная система в идеале охватывает все информационные процессы целого предприятия, достигая полной согласованности, безызыточности и прозрачности информационных процессов. Такие системы иногда называют системами комплексной автоматизации предприятия.

Практически все ИС сегодня являются одновременно и **информационно-вычислительными (ИВС)**, так как в их состав входят вычислительные машины.

Именно компьютеризация придала информационным системам их современный облик, на несколько порядков повысила эффективность и расширила сферу их применения.

Примеры компьютерных информационных систем:

- информационное моделирование,
- базы данных и СУБД, языки управления реляционными базами данных,
- система искусственного интеллекта,
- часть функций локальных и глобальных сетей.

Анализ содержания и систематизация функций ИВС, управляющей крупным объектом (корпорацией, фирмой), позволили выделить и определить следующие обобщенные **функции** ИВС:

- **вычислительная**: своевременная и качественная обработка данных;
- **коммуникационная** - оперативная передача информации в заданные пункты;
- **информирующая** обеспечение быстрого доступа, поиска и выдачи необходимой информации;
- **архивирующая** — выполнение непрерывного накопления, систематизации, хранения и обновления всей необходимой информации;
- **регулирующая** - осуществление информационно-управляющего воздействия на объект управления при отклонении параметров его функционирования от запланированных значений;
- **оптимизирующая** - обеспечение оптимальных плановых расчетов по мере изменения целей, критериев и условий функционирования объекта;
- **самоорганизующаяся** - гибкое изменение структуры и параметров ИВС для достижения вновь поставленных целей;
- **самосовершенствующаяся** - накопление и анализ опыта с целью обоснованного отбора лучших методов проектирования, производства и управления;
- **исследовательская** — обеспечение выполнения научных исследований корпоративных проблем, процессов создания новой техники и технологий, формирования тематики целевых программ комплексных научных исследований;
- **прогнозирующая** выявление основных тенденций, закономерностей и показателей развития объекта и окружающей среды;
- **анализирующая** - определение основных показателей деятельности объекта;
- **синтезирующая** — обеспечение автоматизированной разработки нормативов технологической, финансовой и хозяйственной деятельности;
- **контролирующая** - выполнение автоматизированного контроля качества средств производства, выпускае-

мой продукции;

- *диагностическая* - выполнение автоматизированных процедур диагностики состояния объекта управления;
- *документирующая* — обеспечение формирования необходимых учетно-расчетных, планово-распорядительных, финансовых и других форм документов.

3. Структура ИС ([1])

Для реализации перечисленных функций ИС должна включать набор подсистем. Функциональные подсистемы ИС предназначены для реализации и поддержки моделей, методов и алгоритмов получения управляющей информации.

Практически все рассмотренные разновидности ИС независимо от сферы их применения включают один и тот же набор компонентов:

Структура:

1. Функциональные компоненты.

Функция управления - специальная постоянная обязанность) одного или нескольких лиц, выполнение которых приводит к достижению определенного результата.

Система функций управления – полный комплекс взаимосвязанных во времени и пространстве работ по управлению, необходимых для достижения поставленных перед предприятием целей.

- 1.1. Функциональные подсистемы (модули, бизнес-приложения)
- 1.2. Функциональные задачи,
- 1.3. Модели и алгоритмы

Состав функциональных подсистем зависит от предметной области использования ИС. Каждая из подсистем обеспечивает выполнение комплексов задач и процедур обработки информации, необходимых для эффективного управления объектом.

Например, **состав функциональных подсистем для производственных организаций:**

- **Научно-техническая подготовка производства** отвечает за выполнение научно-исследовательских работ, конструкторскую и технологическую подготовку производства.
- **Бизнес-планирование** выполняет технико-экономическое и оперативно-календарное планирование производства, обеспечивает формирование бизнес-планов.
- **Оперативное управление** - предназначена для управления ходом производства, а также выполняет управление материальными потоками, снабжением и сбытом, учетом затрат на производство.
- **Финансовый менеджмент** - отвечает за формирование финансового плана и портфеля заказов предприятия, анализ результатов его хозяйственной деятельности.
- **Бухгалтерский учет** обеспечивает составление отчетности и учет труда и заработной платы, товарно-материальных ценностей, основных средств, результатов финансовых операций.

2. Обеспечивающие компоненты

2.1. Компоненты системы обработки данных.

- **Информационное обеспечение** - представляет собой совокупность реализованных решений по объемам, размещению и формам организации информации, передаваемой в системе управления. Информационное обеспечение это методы и средства организации информационной базы системы, которые включают системы классификации и кодирования информации, унифицированные системы документов, схемы информационных потоков, методики построения баз данных.
- **Программное обеспечение** - включает в себя совокупность программ регулярного Применения, необходимых для решения функциональных задач и программ, позволяющих наиболее эффективно использовать вычислительную технику, обеспечивая пользователям наибольшие удобства в работе.
- **Техническое обеспечение** - представляет собой комплекс технических средств, задействованных в технологическом процессе преобразования информации в системе. В первую очередь это вычислительные машины, периферийное оборудование, аппаратура и каналы передачи данных.
- **Математическое обеспечение** - совокупность математических методов, моделей и алгоритмов обработки информации, используемых в системе.
- **Правовое обеспечение** - совокупность правовых норм, регламентирующих создание и функционирование информационной системы, порядок получения, преобразования и использования информации.
- **Лингвистическое обеспечение** - совокупность языковых средств, используемых в системе с це-

люю повышения качества ее разработки и облегчения общения человека с машиной.

2.2. Организационные компоненты.

- *Кадровое обеспечение* (персонал) - состав специалистов, участвующих в создании и работе системы, штатное расписание и функциональные обязанности, должностные инструкции
- *Эргономическое обеспечение* - совокупность методов и средств, используемых при разработке и функционировании ИС, создающих оптимальные условия для деятельности персонала, для быстрого освоения системы.
- *Организационное обеспечение* - представляет собой комплекс решений, регламентирующих процессы создания и функционирования как системы в целом, так и ее персонала.



Схема не совсем коорректная

Созданию ИС предшествует исследование предметной области и построение модели автоматизируемого объекта — предприятия.

Разработаны десятки **методологий построения формализованных моделей функционирования предприятия**. Их можно разделить:

1. *Структурные методологии.*
2. *Объектно-ориентированные.*

Структурные методы имеют наибольшее распространение. Структурным принято называть такой метод исследования системы или процесса, который начинается с общего обзора объекта исследования, а затем предполагает его последовательную детализацию. Структурные методы имеют три основные особенности:

- расчленение сложной системы на части, представляемые как «черные ящики», каждый из них выполняет определенную функцию системы управления;
- иерархическое упорядочение выделенных элементов системы с определением взаимосвязей между ними;
- использование графического представления взаимосвязей элементов системы.

Модель, построенная с применением структурных методов, представляет собой иерархический набор диаграмм, графически изображающих выполняемые системой функции и взаимосвязи между ними. Попросту говоря, это рисунки, на которых показан набор прямоугольников, определенным образом связанных между собой. В диаграммы также включается текстовая информация для обеспечения точного определения содержания функций и взаимосвязей. Использование графического представления повышает наглядность модели и облегчает процесс ее восприятия.

Примеры методологий структурного анализа:

- ♦ SADT (Structured Analysis and Design Technique) технология структурного анализа и проектирования, ее подмножество — стандарт IDEF0;
- ♦ DFD (Data Flow Diagrams) - диаграммы потоков данных;
- ♦ ERD (Entity-Relationship Diagrams) - диаграммы «сущность—отношение»
- ♦ STD (State Transition Diagrams) — диаграммы переходов состояний.

Объектно-ориентированный подход к построению моделей информационных систем отличается от структурного большим уровнем абстракции и основывается на представлении системы в виде совокупности объектов, взаимодействующих между собой путем передачи определенных сообщений. В качестве объемов предметной области могут служить конкретные предметы или абстрагированные сущности. Следует отметить, что объектно-ориентированный подход не противопоставляется структурному, а может служить его дополнением.

Жизненный цикл ИС (см.создание АИС) ([1], п.12.3)

Необходимость проектирования ИС может обуславливаться разработкой и внедрением информационных технологий в организации (построение новой информационной системы) либо при модернизации существующих информационных процессов, либо при реорганизации деятельности предприятия (проведении бизнес-реинжиниринга).

Потребности проектирования ИС указывают:

1. для достижения каких целей необходимо разработать систему;
2. к какому моменту времени целесообразно осуществить разработку;
3. какие затраты необходимо осуществить для проектирования системы.

Проектирование ИС является трудоемким, длительным и динамическим процессом. Технологии проектирования, применяемые в современных условиях, предполагают поэтапную разработку системы. Этапы по общности целей могут объединяться в стадии.

Совокупность стадий и этапов, которые проходит ИС в своем развитии от момента принятия решения о создании системы до момента прекращения функционирования системы, называется *жизненным циклом ИС*.

Содержание жизненного цикла разработки ИС сводится к выполнению следующих стадий:

- 1) **Планирование и анализ требований** (предпроектная стадия) — системный анализ. Проводится исследование и анализ существующей информационной системы, определяются требования к создаваемой ИС, формируются технико-экономическое обоснование (ТЭО) и техническое задание (ТЗ) на разработку ИС.

Системный анализ. Основными целями этапа являются: 1) формулировка потребностей в новой ИС (определение всех недостатков существующей ИС); 2) выбор направления и определение экономической обоснованности проектирования ИС.

Системный анализ ИС начинается с описания и анализа функционирования рассматриваемого объекта в соответствии с требованиями (целями), которые предъявляются к нему. В результате этого этапа выявляются недостатки существующей ИС, на основе которых формулируется потребность в совершенствовании системы управления этим объектом, и ставится задача определения экономически обоснованной необходимости автоматизации определенных функций управления (создается технико-экономическое обоснование проекта ИС).

После определения этой потребности возникает проблема выбора направлений совершенствования объекта на основе выбора программно-технических средств. Результаты оформляются в виде технического задания на проект, в котором отражаются технические условия и требования к ИС, а также ограничения по ресурсам проектирования. Требования к ИС определяются в терминах функций, реализуемых системой.

- 2) **Проектирование (техническое и логическое проектирование).** В соответствии с требованиями формируются состав автоматизируемых функций (функциональная архитектура) и состав обеспечивающих подсистем (системная архитектура), проводится оформление технического проекта ИС;

Этап проектирования предполагает:

- 1) проектирование функциональной архитектуры ИС, которая отражает структуру функциональных подсистем и связей между ними;
- является наиболее ответственным и важным этапом точки зрения качества всей последующей разработки ИС

2) проектирование системной архитектуры ИС (состав обеспечивающих подсистем);

Построение системной архитектуры на основе функциональной предполагает определение элементов и модулей информационного, технического, программного обеспечения и других обеспечивающих подсистем, связей по информации и управлению между выделенными элементами и разработку технологии обработки информации.

3) **Реализация** (рабочее и физическое проектирование, кодирование). Разработка и настройка программ, формулировка рабочих инструкций для персонала, создание информационного обеспечения, включая формирование и наполнение баз данных, оформление рабочего проекта.

4) **Внедрение (опытная эксплуатация)**. Комплексная отладка подсистем ИС, т.е. проверка работоспособности элементов и модулей проекта, устранение ошибок на уровне элементов и связей между ними; обучение персонала; поэтапное внедрение ПС в эксплуатацию по подразделениям организации, оформление акта о приемо-сдаточных испытаниях ИС;

5) **Эксплуатация ИС (сопровождение, модернизация)**. Проверка проекта на уровне функций, контроля соответствия его требованиям, сформулированным на стадии системного анализа Сбор статистики о функционировании ИС. исправление недоработок и ошибок, оформление требований к модернизации ИС и ее выполнение (повторение стадии 2 -4).

Важной особенностью жизненного цикла ИС является его повторяемость (цикличность) «системный анализ» - «разработка сопровождение — системный анализ». Это соответствует представлению об ИС, как о развивающейся, динамической системе.

При первом выполнении стадии «Разработка» создается проект ПС, а при последующих реализациях данной стадии осуществляется модификация проекта для поддержания его в актуальном состоянии.

Существуют различные модели жизненного цикла. Среди известных можно выделить следующие:

1. **Каскадная модель** (до 70-х годов) - последовательный переход на следующий этап только после полного после завершения предыдущего;

Достоинство - планирование времени осуществления всех этапов проекта, упорядочении хода конструирования.

Недостатки каскадной модели:

♦ модель недостаточно гибкая - реальные проекты часто требуют отклонения от стандартной последовательности шагов;

♦ цикл основан на точной формулировке исходных требований к ПО (реально в начале проекта требования заказчика определены лишь частично);

♦ результаты проекта доступны заказчику только в конце работы.

2. **Итерационная модель** (70-80-е годы) — с итерационными возвратами на предыдущие этапы после выполнения очередного этапа;

Построение комплексных ИС подразумевает согласование проектных решений, получаемых при реализации отдельных задач. Подход к проектированию «снизу вверх» предполагает необходимость таких итерационных возвратов, когда проектные решения по отдельным задачам объединяются и общие системные решения, и при этом возникает потребность в пересмотре ранее сформулированных требований.

Недостаток: Вследствие большого числа итераций возникают рассогласования и несоответствия в выполненных проектных решениях и документации.

3. **Спиральная модель** (80- 90-е годы) прототипная модель, предполагающая постепенное расширение ПО.

Спиральная модель определяет четыре действия, представляемые четырьмя квадрантами спирали:

♦ планирование - определение целей, вариантов и ограничений;

♦ анализ риска - анализ вариантов и распознавание (выбор) риска;

♦ конструирование — разработка продукта следующего уровня;

♦ оценивание оценка заказчиком текущих результатов конструирования.

С каждой итерацией по спирали (продвижением от центра к периферии) строятся все более полные версии ПО.

Спиральная модель жизненного цикла ИС реально отображает разработку программного обеспечения; позволяет явно учитывать риск на каждом витке эволюции разработки; включает шаг системного подхода в итерационную структуру разработки; использует моделирование для уменьшения риска и совершенствования программного изделия.

Недостатками спиральной модели являются:

♦ новизна (отсутствует достаточная статистика эффективности модели):

- ♦ повышенные требования к заказчику;
- ♦ трудности контроля и управления временем разработки.

В основе спиральной модели жизненного цикла лежит применение **прототипной технологии или RAD-технологии (rapid application development — технологии быстрой разработки приложений)**. Основная идея этой технологии заключается в том, что ИС разрабатывается путем расширения программных прототипов, повторяя путь от детализации требований к детализации программного кода.

При прототипной технологии сокращается число итераций, возникает меньше ошибок и несоответствий, которые необходимо исправлять на последующих итерациях, а само проектирование ИС осуществляется более быстрыми темпами, упрощается создание проектной документации. Для более точного соответствия проектной документации разработанной ИС все большее значение придается использованию CASE-технологий.

RAD-технология обеспечивает экстремально короткий цикл разработки ИС. При полностью определенных требованиях и ограниченной проектной области RAD-технология позволяет создать полностью функциональную систему за очень короткое время (60-90 дней). Выделяют следующие этапы разработки ИС с использованием RAD-технологии:

1) бизнес-моделирование. Моделируется информационный поток между бизнес-функциями. Определяются ответы на вопросы: Какая информация руководит бизнес-процессом? Какая информация генерируется? Кто генерирует ее? Где информация применяется? Кто обрабатывает информацию?

2) моделирование данных. Информационный поток отображается в набор объектов данных, которые требуются для поддержки деятельности организации. Определяются характеристики (свойства, атрибуты) каждого объекта, отношения между объектами;

3) моделирование обработки. Определяются преобразования объектов данных, обеспечивающие реализацию бизнес-функций. Создаются описания обработки для добавления, модификации, удаления или нахождения (исправления) объектов данных;

4) генерация приложения. Предполагается использование методов, ориентированных на языки программирования 4-го поколения. Вместо создания ПО с помощью языков программирования 3-го поколения, RAD-процесс работает с повторно используемыми программными компонентами или создает повторно используемые компоненты. Для обеспечения конструирования используются утилиты автоматизации (CASE-средства);

5) тестирование и объединение. Поскольку применяются повторно используемые компоненты, многие программные элементы уже протестированы, что сокращает время тестирования (хотя все новые элементы должны быть протестированы).

Применение RAD имеет и свои недостатки, и ограничения:

- ♦ большие проекты в RAD требуют существенных людских ресурсов (необходимо создать достаточное количество групп);
- ♦ RAD применима только для приложений, которые можно разделять на отдельные модули и в которых производительность не является критической величиной;
- ♦ RAD неприменима в условиях ВЫСОКИХ технических рисков.

4. Основы применения инструментальных средств ИТ ([1], п.12.4)

Переход на промышленную технологию производства программ, стремление к сокращению сроков, трудовых и материальных затрат на производство и эксплуатацию программ, обеспечение гарантированного уровня качества ИС обусловили бурно развивающееся направление — программотехнику, связанное с технологией создания программных продуктов.

Инструментарий технологии программирования — программные продукты поддержки технологии программирования.

В рамках этих направлений сформировались следующие группы:

- ♦ средства для создания приложений;
- ♦ CASE-технологии (Computer-Aided Software Engineering), предназначенные для автоматизации процессом разработки и реализации информационных систем.

Средства для создания приложений включают языки и системы программирования, а также инструментальную среду разработчика.

Язык программирования - формализованный язык для описания алгоритма решения задачи на компьютере.

Средства для создания приложений - совокупность языков и систем программирования, а также различные программные комплексы для отладки и поддержки разрабатываемых программных продуктов.

Языки программирования разделяют на следующие классы (по синтаксису конструкций языка):

- ♦ машинные языки - языки программирования, воспринимаемые аппаратной частью компьютера (машинные коды);
- ♦ машинно-ориентированные языки - языки программирования, которые отражают структуру конкретного типа компьютера (ассемблеры);

- ♦ алгоритмические языки, не зависящие от архитектуры компьютера языки программирования для отражения структуры алгоритма (Паскаль, Фортран, Бейсик и др.);
- ♦ процедурно-ориентированные языки - языки программирования, где имеется возможность описания программы как совокупности процедур, подпрограмм:
- ♦ проблемно-ориентированные языки - языки программирования, предназначенные для решения задач определенного класса (ЛИСП, РПГ, Симула и др.);
- ♦ интегрированные системы программирования.

Другой классификацией языков программирования является их деление на языки, предназначенные для реализации основ структурного программирования, и объектно-ориентированные языки, поддерживающие понятие объектов, их свойств и методов обработки.

Программа, написанная на языке программирования, проходит этап трансляции, когда происходит преобразование исходного кода программы и объектный код, который далее пригоден к обработке редактором связей. Редактор связей - специальная подпрограмма, обеспечивающая построение загрузочного модуля, пригодного к выполнению.

Трансляция может выполняться с использованием средств компиляторов или интерпретаторов. Компиляторы транслируют всю программу, но без ее выполнения. Интерпретаторы, в отличие от компиляторов, выполняют пооператорную обработку и выполнение программы.

Необходимым средством для профессионального разработчика являются специальные программы, предназначенные для трассировки и анализа выполнения других программ, отладчики.

Современная система программирования состоит из следующих компонентов¹:

- ♦ компилятор;
- ♦ интегрированная среда разработчика программ;
- ♦ отладчик;
- ♦ средства оптимизации кода программ;
- ♦ набор библиотек (возможно, с исходными текстами программ);
- ♦ редактор связей;
- ♦ сервисные средства (утилиты) для работы с библиотеками, текстовыми и двоичными файлами;
- ♦ справочные системы;
- ♦ документатор исходного кода программы;
- ♦ систему поддержки и управления проектом программного комплекса.

CASE-технология — программный комплекс, автоматизирующий весь технологический процесс анализа, проектирования, разработки и сопровождения сложных программных систем.

Основное преимущество CASE-технологии — возможность коллективной работы над проектом за счет поддержки работы разработчиков в локальной сети, экспорта-импорта любых фрагментов проекта организационного управления проектом.

5. Создание автоматизированных информационных систем ([1], п.12.5)

Создание автоматизированных информационных систем регламентируется комплексом стандартов и руководящих документов. Можно выделить следующие стадии и этапы создания АИС (табл.).

Таблица. Стадии и этапы создания АИС

Стадии	Этапы работ
1. Формирование требований к ИС	1.1.Обследование объекта и обоснование необходимости создания ИС. 1.2. Формирование требований пользователя к ИС. 1.3. Оформление отчета о выполненной работе и заявки на разработку ИС (тактико-технического задания)
2. Разработка концепции ИС	2.1. Изучение объекта. 2.2. Проведение необходимых научно исследовательских работ. 2.3. Разработка вариантов концепции ИС и выбор варианта концепции ИС, удовлетворяющего требованиям пользователя. 2.4. Оформление отчета о выполненной работе
3. Техническое задание	3.1. Разработка и утверждение технического задания на создание ИС (см. ниже)
4. Эскизный проект	4.1. Разработка предварительных проектных решений по системе и ее частям. (см. ниже) 4.2. Разработка документации на ИС и ее части
5. Технический проект	5.1. Разработка проектных решений по системе и ее частям. (см. ниже) 5.2. Разработка документации на ИС и ее части.

ИНФОРМАЦИОННЫЕ СИСТЕМЫ (расширенная версия)

	5.3. Разработка и оформление документации на поставку изделий для комплектования ИС и (или) технических требований (технических заданий) на их разработку. 5.4. Разработка заданий на проектирование в смежных частях проекта объекта автоматизации
6. Рабочая документация	6.1. Разработка рабочей документации на систему и ее части. 6.2. Разработка или адаптация программ (см. ниже)
7. Ввод в действие	7.1. Подготовка объекта автоматизации к вводу ИС в действие. 7.2. Подготовка персонала. 7.3. Комплектация ИС поставляемыми изделиями (программными и техническими средствами, программно-техническими комплексами, информационными изделиями). 7.4. Строительно-монтажные работы. 7.5. Пусконаладочные работы. 7.6. Проведение предварительных испытаний. 7.7. Проведение опытной эксплуатации. 7.8. Проведение приемочных испытаний

Одним из центральных элементов всего процесса создания АИС является разработка **технического задания**, структура которого, согласно ГОСТ 34.602-89, содержит следующие разделы:

1. общие сведения;
2. назначение и цели создания (развития) системы;
3. характеристика объектов автоматизации;
4. требования к системе:
 - 4.1. требования к системе в целом - отражают концептуальные параметры и характеристики создаваемой системы, среди которых указываются требования к структуре и функционированию системы, к надежности и безопасности, к численности и квалификации персонала и т. д.
 - 4.2. требования к функциям (задачам), выполняемым системой - содержат перечень функций, задач или их комплексов; временной регламент каждой функции, задачи или комплекса задач; требования к качеству реализации каждой функции; к форме представления выходной информации; характеристики необходимой точности и времени выполнения, требования одновременности выполнения группы функций; достоверности выдачи результатов.
 - 4.3. требования к видам обеспечения:
 - к составу, структуре и способам организации данных в системе (информационно-логическая схема);
 - к информационному обмену между компонентами системы;
 - к информационной совместимости со смежными системами;
 - по использованию общероссийских и других классификаторов, унифицированных документов;
 - по применению систем управления базами данных;
 - к структуре процесса сбора, обработки, передачи данных в системе и представлению данных;
 - к защите данных от разрушений при авариях и сбоях в электропитании системы;
 - к контролю, хранению, обновлению и восстановлению данных;
 - к процедуре придания юридической силы документам, производимым техническими средствами ИС.
5. состав и содержание работ по созданию системы;
6. порядок контроля и приемки системы;
7. требования к составу и содержанию работ ПО подготовке объекта автоматизации к вводу системы в действие;
8. требования к документированию;
9. источники разработки.

На основе установленных в техническом задании основных требований и технических решений на последующих этапах конкретизируются и непосредственно разрабатываются компоненты и элементы системы.

В частности, на этапе 4.1 «Разработка предварительных проектных решений по системе и ее частям» определяются:

- ♦ функции ИС;
- ♦ функции подсистем;
- ♦ концепция информационной базы и ее укрупненная структура;
- ♦ функции системы управления базой данных;
- ♦ состав вычислительной системы;
- ♦ функции и параметры основных программных средств.

На этапе 5.1 «Разработка проектных решений по системе и ее частям» осуществляется разработка об-

щих решений по системе и ее частям:

- ♦ по функционально-алгоритмической структуре системы;
- ♦ по функциям персонала и организационной структуре;
- ♦ по структуре технических средств;
- ♦ по алгоритмам решения задач и применяемым языкам;
- ♦ по организации и ведению информационной базы (структура базы данных);
- ♦ по системе классификации и кодирования информации (словарно-классификационная база);
- ♦ по программному обеспечению.

Разработка и документация программного обеспечения в процессе создания или комплектования автоматизированных систем (**п. 6.2**) регламентируются комплексом стандартов, объединенных в группу «Единая система программной документации (ЕСПД)».

6. Применение информационных систем

6.1. В науке (обзорно)

6.1.1. Автоматизированные системы научных исследований (АСНИ) (По классификации 2.1.).

Автоматизированные системы научных исследований (АСНИ) предназначены для автоматизации научных экспериментов, а также для осуществления моделирования исследуемых объектов, явлений и процессов, изучение которых традиционными средствами затруднено или невозможно.

В настоящее время научные исследования во многих областях знаний проводят большие коллективы ученых, инженеров и конструкторов с помощью весьма сложного и дорогого оборудования.

Большие затраты ресурсов для проведения исследований обусловили необходимость повышения эффективности всей работы.

Эффективность научных исследований в значительной степени связана с уровнем использования компьютерной техники.

Компьютеры в АСНИ используются в информационно-поисковых и экспертных системах, а также решают следующие задачи:

- управление экспериментом;
- подготовка отчетов и документации;
- поддержание базы экспериментальных данных и др.

В результате применения АСНИ возникают следующие положительные моменты:

- в несколько раз сокращается время проведения исследования;
- увеличивается точность и достоверность результатов;
- усиливается контроль за ходом эксперимента;
- сокращается количество участников эксперимента;
- повышается качество и информативность эксперимента за счет увеличения числа контролируемых параметров и более тщательной обработки данных;
- результаты экспериментов выводятся оперативно в наиболее удобной форме — графической или символьной (например, значения функции многих переменных выводятся средствами машинной графики). На экране одного графического монитора возможно формирование целой системы приборных шкал (вольтметров, амперметров и др.), регистрирующих параметры экспериментального объекта.

6.1.2. Системы автоматизированного проектирования (САПР) (По классификации 2.2.).

Системы автоматизированного проектирования (САПР) предназначены для выполнения проектных работ с применением математических методов и компьютерной техники.

Системы САПР широко используются в архитектуре, электронике, энергетике, механике и др. В процессе автоматизированного проектирования в качестве входной информации используются технические знания специалистов, которые вводят проектные требования, уточняют результаты, проверяют полученную конструкцию, изменяют ее и т.д.

Кроме того, в САПР накапливается информация, поступающая из библиотек стандартов (данные о типовых элементах конструкций, их размерах, стоимости и др.).

В процессе проектирования разработчик вызывает определенные программы и выполняет их. Из САПР информация выдается в виде готовых комплектов законченной технической и проектной документации. Такие ИС помогают осуществить:

- Разработку новых изделий и технологий их производства,
- Различные инженерные расчеты (определение технических параметров изделий, расходных норм – трудовых, материальных и т.д.).
- Создание графической документации (чертежей, схем, планировок),
- Моделирование проектируемых объектов,
- Создание управляющих программ для станков с числовым программным управлением (ЧПУ).

6.1.3. Какая взаимосвязь между АСНИ и САПР?

Каждая из систем АСНИ и САПР, конечно, имеет свою специфику и отличается поставленными целями и методами их достижения. Однако очень часто между обоими типами систем обнаруживается тесная связь.

Например, в процессе проектирования может потребоваться выполнение того или иного исследования, и, наоборот, в ходе научного исследования может возникнуть потребность и в конструировании нового прибора и в проектировании научного эксперимента.

Такая взаимосвязь приводит к тому, что на самом деле «чистых» АСНИ и САПР не бывает: в каждой из них можно найти общие элементы. С повышением их интеллектуальности они сближаются. В конечном счете и те и другие должны представлять собой экспертную систему, ориентированную на решение задач конкретной области.

6.2. В обучении (обзорно)

6.2.1. Автоматизированные обучающие системы (АОС) (По классификации 2.5.).

Процесс подготовки квалифицированных специалистов длителен и сложен. Обучение в средней школе и затем в вузе занимает почти треть продолжительности жизни человека. К тому же *в современном информационном обществе знания очень быстро стареют*. Чтобы быть способным выполнять ту или иную профессиональную деятельность, *специалисту необходимо непрерывно пополнять своё образование*. В информационном обществе знать "КАК" важнее чем знать "ЧТО".

Поэтому в наше время основная задача среднего и высшего этапов образования состоит не в том, чтобы сообщить как можно больший объем знаний, а в том, чтобы научить эти знания добывать самостоятельно и творчески применять для получения нового знания. Реально это возможно лишь с введением в образовательный процесс средств новых информационных технологий (СНИТ), ориентированных на реализацию целей обучения и воспитания.

Одним из перспективных направлений *использования СНИТ в образовании* являются Автоматизированные обучающие системы (АОС).

Автоматизированные обучающие системы (АОС) — комплексы программно-технических и учебно-методических средств, обеспечивающих активную учебную деятельность. АОС обеспечивают не только обучение конкретным знаниям, но и проверку ответов учащихся, возможность подсказки, занимательность изучаемого материала и др.

АОС представляют собой сложные человеко-машинные системы, в которых объединяется в одно целое ряд дисциплин: *дидактика* (научно обосновываются цели, содержание, закономерности и принципы обучения); *психология* (учитываются особенности характера и душевный склад обучаемого); *моделирование, машинная графика* и др.

Основное средство взаимодействия обучаемого с АОС — диалог. Диалогом с обучающей системой может управлять как сам обучаемый, так и система. В первом случае обучаемый сам определяет режим своей работы с АОС, выбирая способ изучения материала, который соответствует его индивидуальным способностям. Во втором случае методику и способ изучения материала выбирает система, предъявляя обучаемому в соответствии со сценарием кадры учебного материала и вопросы к ним. Свои ответы обучаемый вводит в систему, которая истолковывает для себя их смысл и выдает сообщение о характере ответа. В зависимости от степени правильности ответа, либо от вопросов обучаемого система организует запуск тех или иных путей сценария обучения, выбирая стратегию обучения и приспособляясь к уровню знаний обучаемого.

Экспертные обучающие системы (ЭОС). Реализуют обучающие функции и содержат знания из определенной достаточно узкой предметной области. ЭОС располагают возможностями пояснения стратегии и тактики решения задачи изучаемой предметной области и обеспечивают контроль уровня знаний, умений и навыков с диагностикой ошибок по результатам обучения.

Учебные базы данных (УБД) и учебные базы знаний (УБЗ), ориентированные на некоторую предметную область. УБД позволяют формировать наборы данных для заданной учебной задачи и осуществлять

выбор, сортировку, анализ и обработку содержащейся в этих наборах информации. В УБЗ, как правило, содержится описание основных понятий предметной области, стратегия и тактика решения задач; комплекс предлагаемых упражнений, примеров и задач предметной области, а также перечень возможных ошибок обучаемого и информация для их исправления; база данных, содержащая перечень методических приемов и организационных форм обучения.

6.3. ИС управления ([1], п.12.2, 12.6) (обзорно)

6.3.1. Основные концепции построения ИС управления (ИСУ) ([1], п.12.2).

Современные концепции построения информационных систем управления представлены следующими методологиями:

1. **MRPII (Manufacturing Resource Planning)** — планирование производственных ресурсов;

Он является наиболее распространенным методом управления производством в мире, разработанный в США и поддерживаемый американским обществом по контролю производства и запасов. Эта методология предлагает ряд способов решения задач управления производством (формирование плана предприятия, планирование продаж, планирование производства, планирование потребностей в материальных ресурсах и производственных мощностей, оперативное управление производством). В основе MRPII лежит иерархии планов.

Планы нижних уровней зависят от планов более высоких уровней план высшего уровня предоставляет входные данные, намечаемые показатели и некоторые ограничения для планов низшего уровня, причем результаты планом низшего уровня оказывают обратное воздействие на планы высшего уровня.

По мере использования стандарта MRPII были выявлены его определённые недостатки, после устранения которых появилась новая методология ERP.

2. **ERP (Enterprise Resource Planning)** — планирование ресурсов предприятия;

Основным отличием данной концепции от предшествующей является ориентация на работу с финансовой информацией, возможность планирования не только производственных, но и иных ресурсов предприятия.

Дополнительно к функциям MRPII в концепции ERP появились следующие: прогнозирование спроса, управление проектами, ведение технологической информации, управление затратами, управление финансами, управление кадрами.

Необходимо отметить, что расширение функционала системы за счет возможности комплексного управления не только материальными, но и другими ресурсами предприятия значительно увеличивает ее стоимость и усложняет работы по внедрению подобных систем.

3. **APS (Advanced Planning and Scheduling)** — расширенное управление производственными графиками;

Особенностью этой концепции является возможность решать такие задачи, как «проталкивание» срочного заказа в производственные графики и распределение заданий с учетом приоритетов и ограничений. В системах, реализующих концепции APS, широко используются современные методы оптимизации (математические, эвристические). В настоящее время концепция APS часто применяется при создании специализированных модулей в ERP-системах.

4. **CSRP (Customer Synchronized Resource Planning)** — планирование ресурсов, синхронизированное с потребностями.

Сущность данной концепции заключается в том, что при планировании и управлении компанией нужно учитывать не только основные производственные и материальные ресурсы предприятия, но и все те, которые обычно рассматриваются как «вспомогательные» или «накладные». Это ресурсы, потребляемые во время маркетинговой деятельности, послепродажного обслуживания проданных товаров, перевалочных и обслуживающих операций, а также внутрицеховые ресурсы, т. е. элементы всего жизненного цикла товара.

Действительно, чтобы правильно управлять стоимостью товара, чтобы понимать, сколько стоит продвижение, производство и обслуживание товара данного типа, нужно учитывать все элементы его функционального жизненного цикла.

Реализация концепции CSRP на конкретном предприятии позволяет управлять всеми деловыми процессами с большей степенью адекватности, чем это было с применением ранее рассмотренных методологий. Например, могут быть учтены возможные вариации спецификации изделия или технологической цепочки, что требуется достаточно часто. При расчете себестоимости можно учесть дополнительные операции по тестированию и административному обслуживанию заказа, а также операции по послепродажному обслуживанию, что практически невозможно в MRP/ERP-системах.

Основные аспекты формализованного представления предметной области при проектировании ИСУ рассмотрим на примере разработки муниципальной информационной системы.

В основе методов создания информационной системы города и области (края, республики) лежит моделирование предметной области.

Под **предметной областью** понимается взаимосвязанная совокупность управляемых объектов организации, субъектов управления, автоматизируемых функций управления и программно-технических средств их реализации.

Муниципальные органы власти			
Объекты управления	Субъекты управления	Автоматизируемые аспекты управления	Программно-технологические средства управления
Муниципальные учреждения	Глава администрации	Документооборот	Подсистемы муниципальных информационных системах (МИС)
Население	Зам. главы администрации	Базы данных	ИС отделов
Отделы	Начальник отделов	Расчеты	СУБД
Подразделения	Начальники учреждений	Отчеты	Электронные правовые системы

Рис. Предметная область Муниципальные органы власти

Формализованное представление предметной области позволяет сократить время и сроки проведения проектных работ и получить более эффективный и качественный проект. Без проведения моделирования предметной области велика вероятность получения некачественной муниципальной информационной системы (МИС), в которой может быть допущено большое количество ошибок в решении стратегических вопросов, приводящих к экономическим потерям и высоким затратам на последующее перепроектирование системы.

Основные требования, предъявляемые к моделям предметной области:

- **Формализованность.**

Для представления моделей используются различные формальные языки моделирования.

Современным подходом к построению моделей анализа и проектирования информационных систем является объектно-ориентированный подход. Он предполагает представление окружающего мира в виде объектов, являющихся экземплярами соответствующих классов. Объектно-ориентированный подход продемонстрировал свою эффективность при построении систем в различных предметных областях и является наиболее популярным в настоящее время. Большинство инструментальных средств, операционных систем и современных языков программирования в той или иной мере являются объектно-Ориентированными.

Промышленным объектно-ориентированным стандартом языка моделирования бизнес-процессов и систем с ориентацией на их дальнейшую реализацию в виде программного обеспечения является **Unified Modeling Language (UML)**. Все представления о модели сложной системы фиксируются в UML в виде специальных графических конструкций диаграмм, что позволяет реализовать второе требование:

- **Наглядность и понятность для заказчиков и разработчиков** основывается на применении графических средств отображения модели (например, с помощью диаграмм).

Например, структурные диаграммы, диаграммы классов, диаграммы реализации, диаграмма компонентов, диаграммы прецедентов, диаграммы состояний, диаграммы деятельности и т.п.

- **Реализуемость** подразумевает наличие средств физической реализации модели предметной области МИС

- **Обеспечение оценки эффективности реализации** ((модели предметной области основывается на определенных методах оценки эффективности и вычислении показателей эффективности)).

Ведущее направление развития ИС

Ведущее направление развития информационных систем - совершенствование их интеллектуальных функций, облегчающих работу с ними, а также их сопровождение и развитие. Такими интеллектуальными функциями, которые могут (но не обязаны) быть у интеллектуальной системы, являются:

- **интерфейс с ИС на естественном языке** (речевой ввод и вывод информации, формирование запросов на естественном языке);
- **поддержка принятия решений** - решения задач на основе информации, имеющейся в информационной

системе;

- **машинное обучение** - модификация своей базы знаний в процессе работы интеллектуальной системы, адаптация к проблемной области. Аналогична человеческой способности «набирать опыт».
- **автоматическое доказательство** (вывод) - способность системы выводить новые знания из старых, находить закономерности в базе знаний.
- **интроспекция** - нахождение противоречий, нестыковок в базе знаний, слежение за правильной организацией знаний.
- **доказательство заключения** - способность системы «объяснить» ход её рассуждений по нахождению решения, причем «по первому требованию».

7. Информационно-поисковые системы (По классификации 3.1.).

7.1. Состав Информационно-поисковые системы

В ИПС выделяют следующие компоненты:

1. **массив документов (текстов) или фактов**, выступающих в качестве объекта хранения и поиска;
2. **информационно-поисковый язык**, предназначенный для отображения содержания документов и операций над ними, в том числе и запросов для поиска документов;
3. **правила, алгоритмы, методы индексирования и поиска документов**, позволяющие описывать документы и операции над ними на информационно-поисковом языке;
4. **комплекс программных и аппаратных средств**, с помощью которых реализуются процессы накопления, хранения и поиска документов;
5. **обслуживающий персонал**, включающий администратора банка документов, системных аналитиков, программистов и индексаторов.

7.2. БАНКИ (БАЗЫ) ИНФОРМАЦИИ

Важнейшими подсистемами информационных систем являются *банки информации*.

Классификация банков информации.

К настоящему времени сложились следующие **три основных типа банков информации**:

1. **БАНК ДАННЫХ** - наиболее характерный пример информационной системы.
 В банке данных хранится достаточно универсальная, необходимая для решения разнообразных прикладных задач, информация об определенной предметной области в специальном представлении, чаще всего предполагающем хранение и обработку с помощью компьютеров.
 При этом сами данные образуют базу данных, а банк, наряду с базой, содержит программные средства обработки данных и реализации запросов, т.е. систему управления базой данных (СУБД).
 Итак, основа банка данных - база данных. Определение базы данных, основные понятия, связанные с различными моделями данных (иерархическая, сетевая, реляционная, постреляционная, многомерная, объектно-ориентированная) уже обсуждались ранее.
 Как правило, банки данных являются системами коллективного пользования. К информации, хранимой в них, часто можно получить доступ по телекоммуникационным сетям.
 В современном мире существует огромное число банков данных. В них содержатся сведения коммерческого характера, данные по библиотечным фондам, системам здравоохранения, транспорта и т.д. Классификация банка данных (и ИС их использующей):
 - 1.1. **Фактографические** – элементом данных является запись; (фактография - описание фактов без их анализа, обобщения, освещения),
 - 1.2. **Документальные** – элементом данных является документ (научная статья, приказ, циркуляр, письмо и т.д.). Обычно под документом понимается текстовый файл. Основной задачей документальных информационных систем является хранение и предоставление пользователю документов, содержание которых соответствуют его информационным потребностям.
 - 1.3. **Документально-фактографические** (документы и факты, извлеченные из документов).
2. **БАНК (БАЗА) ЗНАНИЙ** (knowledge base) - совокупность знаний, относящихся к некоторой предметной области или сфере деятельности, формально представленных таким образом, чтобы на их основе можно было осуществлять рассуждения.
 в зависимости от сложности систем, в которых применяются базы знаний, различают:
 - 2.1. БЗ всемирного масштаба – например, Интернет,

- 2.2. БЗ национальные, например, Википедия,
- 2.3. БЗ отраслевые – например, Автомобильная энциклопедия,
- 2.4. БЗ организаций,
- 2.5. БЗ экспертных систем,
- 2.6. БЗ специалистов.

Например, реально существующий Банк Педагогической Информации, созданный в Республиканском институте повышения квалификации работников образования.

С банком работают лица следующих категорий:

- работники образования (учителя, методисты, работники управленческих структур всех уровней, преподаватели вузов, студенты и учащиеся, родители и др.);
- работники информационной системы, подготавливающие информационные и информационно-педагогические модули;
- работники информационной службы (администратор банка, эксперт).

Содержание банка:

- учебная литература; авторские курсы;
- аспекты педагогической науки;
- методы обучения;
- диагностика педагогического профессионализма;
- зарубежная педагогическая информация
- и др.

Тип содержащейся информации и её местонахождения в базе определяются системой поддержки базы знаний. Хорошая поддержка — залог высокой производительности БЗ. Наиболее важный параметр БЗ — качество содержащихся знаний. Лучшие БЗ включают самую релевантную (соответствующую требованиям поиска) и свежую информацию, имеют совершенные системы поиска информации и тщательно продуманную структуру и формат знаний.

Применение баз знаний:

- Простые базы знаний могут использоваться для хранения данных об организации: документации, руководств, статей технического обеспечения.
Главная цель создания таких баз — помочь менее опытным людям найти существующее описание способа решения какой-либо проблемы предметной области.
- Онтология может служить для представления в базе знаний иерархии понятий и их отношений.
- Базы знаний чаще всего используются в контексте экспертных систем, где с их помощью представляются навыки и опыт экспертов, занятых практической деятельностью в соответствующей области (например, в медицине или в математике). Экспертные системы предназначены для построения способа решения специализированных проблем, основываясь на записях БЗ и на пользовательском описании ситуации.
- Искусственный интеллект.

Создание и использование систем искусственного интеллекта потребует огромных баз знаний.

Раздел искусственного интеллекта, изучающий базы знаний и методы работы со знаниями, называется **инженерией знаний**.

3. БАНК ДОКУМЕНТОВ.

Исторически первым типом банков информации явились банки документов или документальные информационно-поисковые системы. Документальные информационно-поисковые системы бурно развивались в 60-е годы, они широко используются в качестве справочного инструмента пользователей научно-технической информацией, в информационном обслуживании управленческих работников, специалистов и др. В настоящее время интерес к этим системам возобновился в связи с развитием глобальных информационных сетей (Internet) и появлением гипертекстовых серверов типа WWW, которые вместе с соответствующими поисковыми системами можно отнести к распределенным банкам документов.

Объектом хранения в таких системах является документ (научная статья, приказ, циркуляр, письмо и т.д.) или факты, извлеченные из документов. Для обеспечения поиска и доступа к таким документам необходима их предварительная семантическая обработка - индексация.

7.3. ИНДЕКСАЦИЯ ДАННЫХ

Важнейший этап обработки нового документа при поступлении его в хранилище документальной информационно-поисковой системы его индексирование.

Индексирование документа складывается из следующих действий:

- выявления основного смыслового содержания документа (с учетом точки зрения автора документа и информационных потребностей пользователя системы);
- описания смыслового содержания документа на информационно-поисковом языке и получения соответствующего поискового образа документа (ПОД).

В основе индексации лежит классификация данных.

7.3.1. КЛАССИФИКАТОРЫ

Классификация – система распределения объектов (любой предмет, процесс, явления и т.п. как материального так и не материального свойства) по группам в соответствии с определенным признаком.

Классификация объектов производится согласно правилам распределения заданного множества объектов на подмножества (классификационные группировки) в соответствии с установленными признаками их различия или сходства. Применяется в Автоматизированных системах управления и обработке информации.

Классифицировать объект – значит указать номер (или наименование класса), к которому относится данный объект.

Классификатор, или классификационная схема (от лат. classis — разряд и facere — делать) — систематизированный перечень наименований объектов, каждому из которых в соответствие дан уникальный код.

Классификатор является стандартным кодовым языком документов, финансовых отчетов и автоматизированных систем.

Виды классификаторов

Классификаторы разрабатываются как на уровне отдельных предприятий (организаций), так и на уровне государств. Существуют следующие уровни классификаторов:

- **международные** — стандартные классификаторы, используемые по всему миру;
- **межгосударственные** — классификаторы, используемые в рамках экономических союзов и других межгосударственных объединений: например, классификаторы используемые в ЕС, СНГ и т. д.
- **национальные, или межотраслевые** — классификаторы, используемые в пределах государства. Не должны противоречить международным классификаторам;
- **отраслевые** — классификаторы, используемые в рамках одной отрасли;
- **системные** — классификаторы, принятые отдельным предприятием (организацией) для применения в рамках своей автоматизированной системы. Они содержат информацию, необходимую для решения задач в конкретной АС и отсутствующую в национальном или отраслевом классификаторе.

Методы классификации

В классификаторах применяется два основных метода классификации: иерархический и фасетный. Выбор между этими двумя методами зависит от особенностей конкретной предметной области. Существуют следующие **требования** для выбранной системы классификации:

- достаточная емкость и необходимая полнота, которые гарантируют охват всех объектов классификации в заданных границах;
- оправданная глубина;
- обеспечение возможности решения комплекса задач различного уровня;
- возможность расширения множества классифицируемых объектов и внесения необходимых изменений в структуры классификации;
- обеспечение возможности сопряжения с другими классификациями однородных объектов;
- обеспечение простоты ведения классификатора.

1. Иерархический метод классификации - метод, при котором заданное множество последовательно делится на подчиненные подмножества, постепенно конкретизируя объект классификации. При этом основанием деления служит некоторый выбранный признак. Совокупность получившихся группировок при этом образует иерархическую древовидную структуру в виде ветвящегося графа, узлами которого являются группировки.

Выбор последовательности признаков зависит, прежде всего, от характера информации. При построении классификации выбор последовательности признаков зависит от вероятности обращения к тому или иному при-

знаку. При этом наиболее вероятным обращения должны соответствовать высшие уровни классификации. Существуют иерархические системы, в которых рубрики включают по 20 и более подклассов, рубрик и подразделов в нисходящем порядке.

Требования к классификатору, построенному на иерархическом методе классификации:

1. Классификационные группировки, расположенные на одной ступени классификатора, не должны пересекаться, то есть не должны включать в себя аналогичных понятий.
2. На каждой ступени классификатора для разделения вышестоящей группировки должен использоваться только один признак.
3. Сумма подмножества всегда должна давать делимое множество объектов; не должна оставаться часть объектов, не вошедших в состав классификационной группировки.

Основными **преимуществами** иерархического метода является большая информационная емкость, традиционность и привычность применения, возможность создания для объектов классификации мнемонических кодов, несущих смысловую нагрузку. (Мнемотехника - совокупность приёмов и способов, облегчающих запоминание и увеличивающих объём памяти путём **образования искусственных ассоциаций**. Замена абстрактных фактов на понятия и представления, имеющие визуальное, аудиальное или кинестетическое представление).

Значительным **недостатком** иерархической классификации является слабая гибкость структуры, обусловленная фиксированным основанием деления и заранее установленным порядком следования, не допускающим включение новых объектов и классификационных группировок. Таким образом, при изменении состава объектов классификации и характеристик с помощью классификационных задач, требуется коренная переработка всей классификационной схемы.

Используется:

- В науке, как метод классификации (например, классификация биологических видов, соответствует общим и частным признакам. Часто этот метод классификации связан с генезисом.
- В социальных институтах, соответствует принципу подчиненности нижних уровней верхним
- При проектировании и эксплуатации технических объектов, соответствует «детализировке» разбиению крупных объектов на более мелкие
- В планировании, как метод детализации планов
- В программировании, как метод порождения от общего предка объектов, обладающих все более детализированными признаками

Примером иерархической классификационной системы является универсальная десятичная классификация (УДК), широко используемая в библиотечном деле и документальных поисковых системах. УДК охватывает весь спектр знаний.

Шифры УДК, которые можно увидеть на обороте титульного листа всех книг, перед заголовками статей в журналах и сборниках, имеют более чем столетнюю историю. В 1905 г. в Брюсселе на французском языке вышло первое сводное издание таблиц десятичной классификации. Эти таблицы были созданы на основе таблиц десятичной классификации Мельвиля Дьюи, впервые изданных в 1876 г.

Каждый класс (первая ступень деления) в УДК содержит группу более или менее близких наук. Например Информатика включена в группу 5 Математика. Естественные науки (в том числе ИТ). Каждая последующая присоединяемая цифра не меняет значения предыдущих, а лишь уточняет их, обозначая более частное понятие. Например, второй разряд выделяет информационные технологии из общего перечня – кодом 004.

Например,

Разделы 1-го уровня

0 Общий отдел

1 Философия. Психология

2 Религия. Теология

3 Общественные науки

5 Математика. Естественные науки (в том числе ИТ)

6 Прикладные науки. Медицина. Техника

7 Искусство. Декоративно-прикладное искусство.

Фотография. Музыка. Игры. Спорт

8 Языкознание. Филология. Художественная литература. Литературоведение

9 География. Биография. История

Основные деления 2-го уровня

0 Общий отдел

00 Общие вопросы науки и культуры

001 Наука в целом. Науковедение. Организация умственного труда

002 Печать в целом. Документация. Научно-техническая информация
 003 Системы письма и письменности. Знаки и символы. Семиотика в целом. Коды. Графическое пред-

ставление мысли

004 Информационные технологии

...

2. Фасетный метод классификации подразумевает параллельное разделение множества объектов на независимые классификационные группировки.

При этом не предполагается жёсткой классификационной структуры и заранее построенных конечных группировок.

Данная система позволяет выбирать признаки классификации независимо друг от друга. Признаки классификации здесь называются **фасетами** (от английского грань, рамка). Каждый фасет содержит совокупность однородных значений определенного классификационного признака. Например, фасет цвет содержит значения: белый, красный, голубой и т. д.

Данная система обычно отображается в виде таблицы, в которой названия столбцов соответствуют определенным фасетам, а в ячейках располагаются их конкретные значения. Сама процедура классификации состоит в присвоении каждому объекту соответствующих значений из фасетов, при этом не обязательно использование всех фасетов.

Группировка фасетов для конкретного объекта может, выражается структурной формулой $K_S = (\Phi_1, \dots, \Phi_i, \Phi_n)$, где K_S – класс, Φ – фасет). При построении данной системы необходимо, чтобы значения, используемые в различных фасетах, не повторялись.

К классификатору, построенному на фасетном методе классификации, предъявляются следующие **требования**:

1. Должен соблюдаться принцип непересекаемости фасета, то есть состав признаков одного фасета не должен повторяться в других фасетах этого же класса;
2. В состав классификатора должны быть включены только такие фасеты и признаки, которые необходимы для решения конкретных задач.

Основным **преимуществом** классификации с использованием фасетного метода является гибкость структуры ее построения. Изменения в любом из фасетов не оказывают существенного влияния на все остальные. Большая гибкость обуславливает хорошую приспособляемость классификации к меняющемуся характеру решаемых задач, для которых она создается. При фасетной классификации появляется возможность поиска по любому сочетанию фасетов.

Недостатками фасетного метода классификации являются неполное использование емкости, нетрадиционность и иногда сложность применения.

Пример 1, Классификация фильмов:

- тип: документальный, игровой, анимация (мультипликация)
- жанр: боевик, комедия, романтика, фантастика
- продолжительность
- год
- страна
- автор
- другие параметры: немой/звуковой, цветной/чёрно-белый, стерео/5.1 и т. п.

Таким образом, каждый фильм находится в категории типа, жанра и современного технического уровня. Так как данные категории независимы, то для каждого конкретного фильма информация будет представлена в виде пересечения данных признаков, которые не исключают друг друга.

Пример 2, Первый разряд фасету «пол», со значениями 1 мужчины, 2 женщины.

Второй разряд фасету «семейное положение» со значением 1 женат, 2 холост.

Третий разряд фасету «возраст», со значениями 1 до 20 лет, 2 от 20 до 30 лет, 3 свыше 30 лет.

Четвертый разряд фасету «название факультета», со значениями 1 МУ, 2 СТ.

Расшифруем код 2 1 1 1 – женщина, замужем, возраст до 20, студентка факультета МУ.

3. Эмпирический (неиерархический) метод классификации

Примером эмпирической (неиерархической) классификации может быть алфавитно-предметная классифи-

кация. Словарный запас такой классификации состоит из упорядоченного по алфавиту множества слов, словосочетаний и фраз естественного языка, обозначающих предметы какой-либо отрасли науки или практической деятельности. Каждому предмету или вопросу при этом отводится только один индекс, собирающий всю информацию относительно данного предмета или вопроса независимо от аспекта рассмотрения.

В предметных классификациях используются следующие термины:

- предметная рубрика (заголовок) - слово или фраза естественного языка, используемая для обозначения основного предмета документа (или запроса);
- предметный подзаголовок - слово или фраза, обозначающая аспект рассмотрения предмета, указанного в предметном заголовке или в подклассе предметов, входящих в класс, обозначенный предметным заголовком;
- предметный словник - упорядоченное по алфавиту множество предметных заголовков, используемых для построения какого-либо каталога или указателя.

Система предметных заголовков и подзаголовков и более мелких разделов создает сложную предметную классификацию по аспектам рассмотрения предмета, т.е. имеет некоторые черты фасетных классификаций.

Методы кодирования в классификаторах

Под **кодированием** понимается присвоение кода классификационной группировке или объекту классификации. Кодирование предназначено для формализованного описания наименований различных аспектов данных. Обычно кодирование представляет собой процесс обозначения исходного множества объектов или сообщений набором символов заданного алфавита на основе совокупности определенных правил.

1. **Последовательный метод** — в кодовом обозначении знаки на каждой ступени деления зависят от результатов разбиения на предыдущих ступенях. В результате кодовое обозначение группировки дает информацию о последовательности признаков, характеризующих эту группировку. Код составляется следующим образом, сначала записывается код 1 уровня, затем 2 и т. д. Наиболее подходит иерархическому методу классификации.
2. **Параллельный метод** — признаки классификации кодируются независимо друг от друга определенными разрядами или группой разрядов кодового обозначения. Метод параллельного кодирования чаще всего используется при фасетной классификации, но применяется также и в иерархической классификации.
3. **Регистрационный метод** применяется для однозначной идентификации объектов путем присвоения уникального кода, не требует предварительной классификации. Используется при эмпирической классификации. Различают:
 - 3.1. **Порядковый метод** — каждый из объектов множества кодируется с помощью текущего номера по порядку. Обеспечивает простоту добавления новых объектов и краткость кода, однако такой код не несёт никакой информации об объекте. Используется в случаях, когда не требуется сложного деления на множества, например, в классификаторе валют.
 - 3.2. **Серийно-порядковый метод** — когда предварительно выделяются группы объектов со сходными признаками, которые образуют серию, а затем уже внутри серии производится порядковая нумерация объектов. Используется для идентификации объектов. Например, классификатор должностей и служащих. Примером этой системы кодирования может служить обычный лотерейный билет. ((МУ08..., № паспорта))

7.3.2. Информационно-поисковые языки (ИПЯ) (+ [1], п.7.1)

Информационно-поисковые языки (ИПЯ) предназначен для отображения содержания документов и операций над ними, в том числе и запросов для поиска документов.

Основными элементами ИПЯ являются:

- **Алфавит ИПЯ** система знаков, используемых для записи слов и выражений ИПЯ.
- **Лексика**, или словарный состав, ИПЯ совокупность слов, словосочетаний и выражений, используемых для построения текстов ИПЯ.
- **Грамматика ИПЯ** совокупность средств и способов построения, изменения и сочетания лексических единиц. Грамматика включает морфологию и синтаксис.

Классификация ИПЯ.

1. **Классификационные языки**, которые, в свою очередь, в соответствии с используемыми в них классификаторами делятся по структуре:
 - 1.1. ИПЯ иерархической структуры;
 - 1.2. ИПЯ фасетной структуры (фасет от фр. *facette*. тех. 1. Скошенная боковая грань чего-н. 2. Грань отшлифованного камня.);
 - 1.3. эмпирические (неиерархические) языки;
2. **Дескрипторные ИПЯ** основаны на методе координатного индексирования.

Сущность которого сводится к тому, что смысловое содержание документа может быть с достаточной точностью и полнотой выражено списком ключевых слов.

Ключевое слово - это лексическая единица информационно-поискового языка, являющаяся любой частью речи кроме союза (существительным, прилагательным, глаголом, наречием, числительным или местоимением) естественного языка или словосочетанием.

Основной критерий отбора ключевых слов из текста - степень их полезности для индексирования документа или запроса.

Координатное индексирование выполняется с помощью ключевых слов и логических операций конъюнкции (логическое "И", логическое умножение) и дизъюнкции (логическое "ИЛИ", включающее "ИЛИ", логическое сложение). Близкие по смыслу ключевые слова образуют классы, имена которых также являются единицами поискового языка и называются **дескрипторами**. Дескрипторы вместе с набором ключевых слов языка и семантических связей между ними образуют **тезаурус** - систематизированный набор данных об области знания, позволяющий в ней ориентироваться – условно словарь-справочник.

В дескрипторной статье тезауруса обычно устанавливаются следующие семантические отношения: отношение синонимии; отношение подчинения; отношение ассоциации.

Такие связи служат для увеличения полноты представления содержания документов и позволяют формировать запрос не обязательно в ключевых словах документа.

Пример дескрипторной статьи ((без таблицы)):

	отношение	
языки алгоритмические	синонимы:	алгоритмические языки
машинно-ориентированные языки проблемно-ориентированные языки	вышестоящие:	программное обеспечение
языки формальные	нижестоящие:	автокоды
алгол паскаль фортран си бейсик	ассоциации:	алгоритмы программирование

3. Комбинированные.

Целесообразность применения того или иного языка во многом зависит от назначения информационной системы, степени ее автоматизации. Для описания документов в библиотеках, общих и технических архивах обычно применяют классификационные языки. В автоматизированных информационно-поисковых системах используются, главным образом, дескрипторные языки.

7.4. Система индексирования ([1], п.7.4)

Система индексирования (СИ) - совокупность методов и средств перевода текстов с естественного языка на ИПЯ в соответствии с заданным набором словарей лексических единиц и с правилами применения ИПЯ.

Классификация систем индексирования:

1. По степени автоматизации процесса индексирования:

- ручного индексирования;
- автоматического индексирования;
- автоматизированного индексирования.

2. По степени контролируемости:

- без словаря;

- с жестким словарем;
- со свободным словарем.

3. По характеру алгоритма отбора слов текста:

- с последовательным просмотром текста (отбираются все полнозначные слова);
- с эвристическими процедурами выбора слов текста (слова отбираются интуитивно или по заданной процедуре):
- со статистическими процедурами выбора слов (отбираются только информативные слова в соответствии с распределением частот их употребления).

4. По характеру лексикографического контроля:

- без лексикографического контроля;
- с полным контролем;
- с промежуточным контролем.

Лексикографический контроль предусматривает:

- ♦ устранение синонимии (совпадение слов по значению, не совпадение по написанию), полисемии (одно и то же слово выражает набор родственных понятий, н-р, слово соль – вещество, а также понятие смысла, например, жизни, похожи по смыслу), и омонимии (совпадение по написанию или звучанию, несовпадение по смыслу) на основе нормативных словарей лексических единиц с парадигматическими отношениями между ними;
- ♦ нормализацию слов на основе морфологических нормативных словарей.

5. По характеру морфологического анализа слов:

- ♦ с использованием морфологических словарей;
 - ♦ с использованием основных лексических словарей;
 - ♦ С использованием морфологического анализа с усечением слов.
- Возможны системы индексирования без морфологического анализа.

Процесс **свободного индексирования** состоит в следующем. Индексатор выписывает слова или словосочетания, которые, по его мнению, отражают содержание текста. Он может брать слова, отсутствующие в тексте, но важные, с его точки зрения, для выражения смысла текста. Отобранный список слов является поисковым образом документа. Это СИ с *ручным индексированием*.

Процесс **полусвободного индексирования** аналогичен вышеописанному, но слова для подберутся только из словаря.

При **жестком индексировании** слова берутся только из текста.

Поначалу индексирование осуществлялось специально подготовленными специалистами-экспертами в предметной области, которые могли осуществлять глубокий анализ смыслового содержания документа и относить его (индексировать) к тем или иным классам, рубрикам, ключевым терминам. В этом случае были высоки накладные расходы, поскольку требовалось наличие в штате высококвалифицированных специалистов-индексаторов. Кроме того, процесс индексирования в некоторой мере был субъективным. Поэтому возникла задача автоматизации индексирования документов.

Существуют два подхода к *автоматическому индексированию*. Первый основан на использовании словаря ключевых слов. Индексирование в таких системах осуществляется путем последовательного автоматического поиска в тексте документа ключевых терминов. Строится индекс, представляющий поисковое пространство документов. Возможны два типа такого индекса - прямой и инвертированный.

Прямой тип индекса строится по схеме «документ—термины». Поисковое пространство в этом случае представлено в виде матрицы размерностью $p \times m$. Строки этой матрицы представляют поисковые образы документов.

Инвертированный тип индекса строится по обратной схеме — «термин—документы». Поисковое пространство соответственно представлено аналогичной матрицей, только в транспонированной форме. Поисковыми образами документов в ЭТОМ случае являются столбцы матрицы.

Прямой индекс

Номер документа	Термины			
	t_1	f_1	t_2	U
d_1	+		+	
d_2	+	+		
d_3		+		+
d_4			f	+

Инвертированный индекс

Термины	Номера документов			
	d	d_2	d_3	d_4
f_i		+		+
g_i	+	+		
h		+	+	
$t <$	+			+
u			+	+

Пример прямого и инвертированного индексов

Второй подход к автоматическому индексированию применяется в *полнотекстовых системах*. В процессе индексирования в индекс заносится информация обо всех словах текста документа (отсюда и название «полнотекстовые»).

7.5. Информационный поиск

Информационный поиск (ИП) (англ. Information retrieval) — процесс поиска неструктурированной документальной информации и наука об этом поиске.

История

Термин «информационный поиск» был впервые введён Кельвином Муром в 1948 в его докторской диссертации, опубликован и употребляется в литературе с 1950.

Сначала системы автоматизированного ИП, или информационно-поисковые системы (ИПС), использовались лишь для управления информационным взрывом в научной литературе.

Термин «информационный взрыв» появился в 50-х гг. и означает постоянное увеличение скорости публикаций). В 1950-1970 годы действительно наблюдался громадный рост научных публикаций. Однако в XXI веке многие научные направления утратили государственную поддержку в таких «научных державах» как Россия и США. В результате количество публикаций резко уменьшилось, а некоторые журналы перестали издаваться.

С другой стороны в XXI веке наметился информационный взрыв в Интернете.

Так например, количество блогов (живые журналы) удваивается каждые 6 месяцев.

А в информационных порталах количество статей растёт экспоненциально.

Многие университеты и публичные библиотеки стали использовать ИПС для обеспечения доступа к книгам, журналам и другим документам. Широкое распространение ИПС получили с появлением сети Интернет. У русскоязычных пользователей наибольшей популярностью пользуются поисковые системы Google, Яндекс и Рамблер.

Информационный поиск

Поиск информации представляет собой процесс выявления в некотором множестве документов (текстов) всех таких, которые посвящены указанной теме (предмету), удовлетворяют заранее определенному условию поиска (запросу) или содержат необходимые (соответствующие информационной потребности) факты, сведения, данные.

Процесс поиска включает последовательность операций, направленных на сбор, обработку и предоставление необходимой информации заинтересованным лицам.

В общем случае поиск информации состоит из четырех этапов:

- определение (уточнение) информационной потребности и формулировка информационного запроса;
- определение совокупности возможных держателей информационных массивов (источников);
- извлечение информации из выявленных информационных массивов;
- ознакомление с полученной информацией и оценка результатов поиска.

Виды поиска

1. **Полнотекстовый поиск** — поиск документа в базе данных текстов на основании содержимого этих документов, а также совокупность методов оптимизации этого процесса. (например, Консультант Плюс, Гарант, Кодекс)

Первые версии программ полнотекстового поиска предполагали сканирование всего содержимого всех документов в поиске заданного слова или фразы. При использовании такой технологии поиск занимал очень много времени (в зависимости от размера базы), а в интернете был бы невыполним.

Современные алгоритмы заранее формируют для поиска так называемый полнотекстовый индекс — словарь, в котором перечислены все слова и указано, в каких местах они встречаются. При наличии такого

индекса достаточно осуществить поиск нужных слов в нём и тогда сразу же будет получен список документов, в которых они встречаются. Пример полнотекстового поиска — любой интернет-поисковик, например www.yandex.ru, www.google.com.

2. **Поиск по метаданным** ((Метаданные это информация о данных))— это поиск по неким атрибутам документа, поддерживаемым системой — название документа, дата создания, размер, автор и т. д. Пример поиска по реквизитам — диалог поиска в файловой системе (например, MS Windows).
3. **Поиск по изображению** — поиск по содержанию изображения. Поисковая система распознает содержание фотографии (загружена пользователем или добавлен URL изображения). В результатах поиска пользователь получает похожие изображения. Так работают поисковые системы: Xcavator, Retrievr, PolarRose, Picollator Online by Recogmission.

Задачи информационного поиска

Центральная задача ИП — помочь пользователю удовлетворить его информационную потребность. Так как описать информационные потребности пользователя технически непросто, они формулируются как некоторый запрос, представляющий из себя набор ключевых слов, характеризующий то, что ищет пользователь.

Классическая задача ИП, с которой началось развитие этой области, — это поиск документов, удовлетворяющих запросу, в рамках некоторой статической коллекции документов.

Но список задач ИП постоянно расширяется и теперь включает:

- **Классификация документов** - отнесение документа к одной из нескольких категорий основываясь на содержании документа.
- **Фильтрация документов** - разбиение множества документов на категории. Однако в отличии от классификации, таких категорий всего две: документы, удовлетворяющие заданному критерию, и не удовлетворяющие ему.
Одной из наиболее важных частных случаев является задача тематической фильтрации документов, то есть автоматического определения документов, соответствующих заданной тематике, за счет отсева прочих документов.;
- **Кластеризация документов** - автоматическое выявление групп семантически похожих документов среди заданного фиксированного множества документов.
Следует отметить, что группы формируются только на основе попарной схожести описаний документов, и никакие характеристики этих групп не задаются заранее, в отличие от классификации документов, где категории задаются заранее.
- Вопросы моделирования;
- Проектирование архитектур поисковых систем и пользовательских интерфейсов
- Извлечение информации, в частности аннотирования и реферирования документов;
- Языки запросов и др.

Запрос и объект запроса

Исходной информацией для поиска является поисковый запрос и объект запроса.

- **Запрос** — это формализованный способ выражения информационных потребностей пользователем системы. Для выражения информационной потребности используется язык поисковых запросов, синтаксис варьируется от системы к системе. Кроме специального языка запросов, современные поисковые системы позволяют вводить запрос на естественном языке.
- **Объект запроса** — это информационная сущность, которая хранится в базе автоматизированной системы поиска. Несмотря на то, что наиболее распространенным объектом запроса является текстовый документ, не существует никаких принципиальных ограничений. В частности, возможен поиск изображений, музыки и другой мультимедиа информации. Процесс занесения объектов поиска в ИПС называется индексацией. Далеко не всегда ИПС хранит точную копию объекта, нередко вместо неё хранится суррогат.

Результатами индексирования документов и запросов являются их поисковые образы.

Формализованное представление индекса документа называется *поисковым образом документа* (ПОД). Пользователь выражает свои информационные потребности посредством специального языка, формируя поисковый образ запроса (ПОЗ) к базе документов. Поиск документов по запросу означает сопоставление поискового образа документа и поискового образа запроса (ПОД и ПОЗ). Качество поиска зависит от критериев смыслового соответствия документа запросу (критериев выдачи).

Соответствие найденных документов запросу пользователя называется *релевантностью*

Оценки эффективности

Существует много способов оценить насколько хорошо документы, найденные ИПС, соответствуют запросу. К сожалению, понятие степени соответствия запроса, или другими словами релевантности, является субъективным понятием, а степень соответствия зависит от конкретного человека, оценивающего результаты выполнения запроса.

Точность (precision). Отношение числа релевантных документов, найденных ИПС, к общему числу найденных документов:

$$\text{Точность} = \frac{D_{\text{релевантных}} \cap D_{\text{найденных}}}{D_{\text{найденных}}} \quad (\cap - \text{пересечение})$$

где $D_{\text{рел}}$ — это множество релевантных документов в базе, а $D_{\text{н}}$ — множество документов, найденных системой.

Полнота (recall). Отношение числа найденных релевантных документов, к общему числу релевантных документов в базе:

$$\text{Полнота} = \frac{D_{\text{релевантных}} \cap D_{\text{найденных}}}{D_{\text{релевантных}}}$$

Выпадение (fall-out). Выпадение характеризует вероятность нахождения нерелевантного ресурса и определяется, как отношение числа найденных нерелевантных документов к общему числу нерелевантных документов в базе:

$$\text{Выпадение} = \frac{D_{\text{нерел}} \cap D_{\text{найденных}}}{D_{\text{нерел}}}$$

где $D_{\text{нерел}}$ — это множество нерелевантных документов в базе, а n — множество документов, найденных системой.

F-мера (F-measure, мера Ван Ризбергера). Традиционно F-мера определяется, как гармоническое среднее точности и полноты:

$$F = \frac{2 \cdot \text{Точность} \cdot \text{Полнота}}{\text{Точность} + \text{Полнота}}$$

8. Поиск в Интернете

«Мировая паутина» WWW (от английских слов World Wide Web - «Всемирная паутина») - это распределенная информационная система мультимедиа, основанная на гипертексте.

Информация хранится в ней на огромном множестве объединенных в сеть серверов - компьютерах с соответствующим программным обеспечением. Пользователи, или клиенты, имеющие доступ к сети, могут получать эту информацию, используя специальное программное обеспечение - программы просмотра документов (так называемые, WWW-браузеры).

Количество информации, размещенной в Интернете, растет с каждым днем, однако уровень ее структуризации низок, а частота обновления весьма высока. Проблема наполнения, стоявшая вначале трансформировалась в проблему поиска.

8.1. Поисковая система

Поисковая система — веб-сайт, предоставляющий возможность поиска информации в Интернете.

Большинство поисковых систем ищут информацию на сайтах Всемирной паутины, но существуют также системы, способные искать файлы на ftp-серверах, товары в интернет-магазинах, а также информацию в группах новостей Usenet.

Обычно составной частью поисковой системы является так называемая поисковая машина.

8.1.2. Поисковая машина

Поисковая машина (поисковый движок) - комплекс программ, обеспечивающий поиска информации в

ПС.

По области поиска поисковые машины можно условно разделить на:

- **Локальные поисковые машины** предназначены для поиска информации по какой-либо части всемирной сети, например по одному или нескольким сайтам, либо по локальной сети.
- **Глобальные поисковые машины** предназначены для поиска информации по всей сети Интернет либо по значительной её части. Представителями таких поисковых машин являются поисковые машины поисковых систем Google, Yahoo и т. д.

Теоретически могут существовать поисковые машины для поиска самой разной информации, например текстов, видео, изображений, людей и др.

При этом документами, по которым осуществляется поиск, могут быть как текстовые форматы (например .html, .htm, .txt, .doc, .rtf, ...), так и форматы для хранения изображений (.gif, .png, .svg, ...), видео, звука и другой информации.

Однако наиболее распространённым на сегодня является поиск по текстовым документам. Такими документами могут быть web-страницы, документы в формате doc, rtf, txt и др.

Поиск по изображениям, видео, звукам более сложен с технологической точки зрения, поэтому массово не реализован. Картинки, как правило ищутся не по самим изображениям, а по альтернативным текстам, соответствующим этим изображениям.

Поисковые машины выполняют **несколько функций**:

- Поиск ссылок на страницы и другие документы сайтов.
Поиск ссылок может осуществляться несколькими путями:
 - 1) Автоматически. При этом поисковая машина ищет ссылки со страниц сайтов.
 - 2) В ручном режиме. Пользователи сами добавляют в базу данных поисковой машины ссылки на страницы своих сайтов.
- Индексация документов сайтов.
Индексация документов сайтов — извлечение из документов информации, важной для поиска, преобразование этой информации в формат, удобный для поисковой машины и сохранение этой информации в базу данных поисковой машины.
- Поиск по базе данных проиндексированных документов.
Поиск по базе данных проиндексированных документов может состоять из нескольких этапов, а именно:
 - 1) Нахождение документов, соответствующих поисковому запросу.
 - 2) Ранжирование документов в соответствии с их релевантностью поисковым запросам.

Основными критериями качества работы поисковой машины являются

- релевантность (степень соответствия запроса и найденного, т.е. уместность результата),
- полнота базы,
- учёт морфологии языка.

Индексация информации в Интернете осуществляется специальными поисковыми роботами.

8.1.3. Поисковый робот

Поисковый робот («веб-паук», краулер) - программа, являющаяся составной частью поисковой системы и предназначенная для обхода страниц Интернета с целью занесения информации о них (ключевые слова) в базу поисковика.

По своей сути паук больше всего напоминает обычный браузер. Он сканирует содержимое страницы, забрасывает его на сервер поисковой машины, которой принадлежит и отправляется по ссылкам на следующие страницы. Владельцы поисковых машин обычно ограничивают глубину проникновения паука внутрь сайта и максимальный размер сканируемого текста, поэтому чересчур большие сайты могут оказаться не полностью проиндексированными поисковой машиной. Кроме обычных пауков, существуют так называемые «**дятлы**» — роботы, которые «простукивают» проиндексированный сайт, чтобы определить, что он подключен к Интернету.

Порядок обхода страниц, частота визитов, защита от заикливания, а также критерии выделения ключевых слов определяется алгоритмами поисковой машины.

В большинстве случаев переход от одной страницы к другой осуществляется по ссылкам, содержащимся на

первой и последующих страницах.

Также многие поисковые системы предоставляют пользователю возможность самостоятельно добавить сайт в очередь для индексирования. Обычно это существенно ускоряет индексирование сайта, а в случаях, когда никакие внешние ссылки не ведут на сайт, вообще оказывается единственной возможностью заявить о его существовании.

Ограничить индексацию сайта можно с помощью файла robots.txt, однако некоторые поисковые системы могут игнорировать наличие этого файла. Полная защита от индексации обеспечивается механизмами, обойти которые пока паукам не под силу. Обычно — установкой пароля на странице, либо требованием заполнить регистрационную форму перед тем, как получить доступ к содержимому страницы.

((robots.txt — файл ограничения доступа к содержимому роботам на http сервере.

Файл должен находиться в корне сайта (т.е. иметь путь относительно имени сайта /robots.txt).

Использование файла добровольно, стандарт был принят консорциумом 30 января 1994 года в списке рассылки robots-request@nexor.co.uk и с тех пор используется большинством известных поисковых машин.

Файл robots.txt используется для частичного управления индексированием сайта поисковыми роботами. Этот файл состоит из набора инструкций для поисковых машин, при помощи которых можно задать области сайта, которые не должны индексироваться.

Файл robots.txt может использоваться для указания расположения файла Sitemaps.

Описание структуры

Файл состоит из записей. Записи разделяются одной или более пустых строк (признак конца строки символы CR, CR+LF, LF). Каждая запись содержит непустые строки следующего вида:

<field>:<optionalspace><value><optionalspace>

где field — это или User-agent или Disallow, а value — значение. optionalspace — необязательный пробел.

8.2. История

Первой поисковой системой для Всемирной паутины был «Wandex», уже не существующий индекс, который создавал «World Wide Web Wanderer» — бот, разработанный Мэтью Грэйем (англ. Matthew Gray) из Массачусетского технологического института в 1993. Также в 1993 году появилась поисковая система «Aliweb», работающая до сих пор. Первой полнотекстовой (т. н. «crawler-based» — то есть индексирующей ресурсы при помощи робота) поисковой системой стала «WebCrawler», запущенная в 1994. В отличие от своих предшественников, она позволяла пользователям искать по любым ключевым словам на любой веб-странице, с тех пор это стало стандартом во всех основных поисковых системах. Кроме того, это был первый поисковик, о котором было известно в широких кругах. В 1994 был запущен «Lycos», разработанный в университете Карнеги Мелона.

Вскоре появилось множество других конкурирующих поисковых машин, таких как «Excite», «Infoseek», «Inktomi», «Northern Light» и «AltaVista». В некотором смысле они конкурировали с популярными интернет-каталогами, такими, как «Yahoo!». Позже каталоги соединились или добавили к себе поисковые машины, чтобы увеличить функциональность. В 1996 году русскоязычным пользователям интернета стало доступно морфологическое расширение к поисковой машине Altavista и оригинальные российские поисковые машины Rambler и Aport.

23 сентября 1997 года была открыта поисковая машина Яндекс.

Хронология		
Год	Система	Событие
1993	Aliweb	Запуск
1994	WebCrawler	Запуск
1994	Lycos	Запуск
1995	AltaVista	Запуск (Создана DEC)
1995	Excite	Запуск
1996	Inktomi	Основана
1996	Ask Jeeves	Основана
1996	Rambler	Запуск
1996	Aport	Запуск
1997	Northern Light	Запуск

1997	Яндекс	Запуск
1998	Google	Запуск
1999	AlltheWeb	Запуск
2000	Teoma	Основана
2004	Yahoo! Search	Окончательный запуск (первые собственные результаты)
2004	MSN Search	Бета запуск
2005	MSN Search	Окончательный запуск
2007	Gogo.ru	Бета запуск
2007	Igde.ru	Запуск
Примечание: годом запуска считается год, когда появилась возможность получать результаты поиска через Всемирную паутину.		

8.3. Популярные поисковые системы

Всеязычные:

Google (32 % Русскоязычного сегмента)

Yahoo! (0,4 % Рунета) и принадлежащие этой компании поисковые машины:

Overture

Inktomi

AltaVista

Alltheweb FAST-Engine

MSN (0,2 % Рунета) (принадлежит компании «Microsoft»)

Англоязычные и международные:

AskJeeves (механизм Teoma)

Русскоязычные — большинство «русскоязычных» поисковых систем индексируют и ищут тексты на многих языках — украинском, белорусском, английском и др. Отличаются же они от «всеязычных» систем, индексирующих все документы подряд, тем, что в основном индексируют ресурсы, расположенные в доменных зонах, где доминирует русский язык или другими способами ограничивают своих роботов русскоязычными сайтами.

Яндекс (44,4 % Рунета)

Rambler (10,6 % Рунета)

Mail.ru (7,3 % Рунета)

Nigma (0,5 % Рунета)

Gogo.ru (0,3 % Рунета)

Aport (0,2 % Рунета)

По данным компании Net Applications в декабре 2007 года рыночная доля Google в мире составляла 77.04 %,

Yahoo — 12.46 %,

MSN — 3.33 %,

Microsoft Live Search — 2.57 %, AOL — 2.12 %, Ask — 1.38 %, AltaVista — 0.13 %, Excite — 0.07 %, Lycos — 0.02 %, All the Web — 0.02 %.

По данным аналитической компании comScore все поисковые сайты в декабре 2007 года обработали 66 млрд 221 млн поисковых запросов.

Яндекс попал в статистику и находится на 9-ом месте.

8.4. Язык поисковых запросов

Как известно, в хорошем вопросе содержится половина ответа. Поэтому важно уметь правильно оставить запрос.

Прежде чем искать, надо четко сформулировать, что надо найти. Постараться детализировать запрос, включив в него как можно больше слов, способных уточнить смысл искомого.

Язык поисковых запросов — набор метасимвол и правил, в соответствии с которыми строится запрос к по-

исковой системе. В различных поисковых системах язык может различаться, однако обычно он является некоторым подобием языка регулярных выражений с дополнениями, связанными со спецификой работы той или иной поисковой системы. Знание и правильное применение языка запросов конкретной поисковой машины улучшает и упрощает пользователю процесс поиска информации.

В языке Perl к Метасимволам относятся следующие символы: "\", ".", "\\", "\$", "[", "(", ")", "*", "+", "?", "}".

Различные метасимволы выполняют в регулярном выражении различные функции, в частности, используются для обозначения одиночного символа или их группы, обозначают привязку к определенному месту строки, число возможных повторений отдельных элементов, возможность выбора из нескольких вариантов и т. д.

Для того чтобы эффективно пользоваться поисковой системой, нужно знать синтаксис запросов – другими словами, те символы, которые служат сигналами для поисковых машин. Для каждой системы они индивидуальны, но общие моменты выделить можно:

1. Если запрос состоит из двух и более слов необходимо поставить знак «+» или «&» перед теми словами, которые определяют суть искомого.
 2. Если надо выбрать какой-либо из двух вариантов, пользуются знаком «|».
 3. Если надо исключить какую-либо область или понятие из искомого текста необходимо поставить знак «-».
- Например, [(стихи | проза) +Набоков - биография] - означает «ищу стихи или прозу Набокова и обхожу вниманием те страницы, где говорится о его биографии»
4. Большинство современных поисковых систем учитывают морфологию слова, т.е. ищут понятия во всех возможных словоформах, не обращая внимания на изменения падежа, лица или числа. Если вас интересуют конкретная форма слова надо поставить оператор «!».
 5. Многие поисковые системы позволяют вести контекстный поиск целой фразы, которую надо предварительно заключить в кавычки.

Язык поисковых запросов состоит из логических операторов, префиксов обязательности, возможности учета расстояния между словами, морфологии языка, регистра слов, расширенных операторов, возможностей расширенного поиска, уточнения поиска. **Сравнительная таблица языка запросов по этим параметрам для четырех поисковых систем рунета.**

**Описание языка запросов наиболее популярных поисковых систем
(сравнительная таблица)**

\	Яндекс	Rambler	Google
Ссылки на описания языков поисковых запросов	Яндекс: Помощь	Rambler: Помощь	Google Help Central
Логические операторы	& - И в пределах предложения, && - И в пределах документа, - ИЛИ, ~ - И-НЕ в пределах предложения, ~~ - И-НЕ в пределах документа, (и) - скобки	AND или & - И, OR или - ИЛИ, NOT или ! - НЕ, (и) - скобки	OR - ИЛИ
Префиксы обязательности	+ и -	Нет	+ и -
Учет морфологии языка (словоформы)	На основе правил, !Слово - указание точной словоформы	Словарный (* и ? - зарезервированы, но пока не используются)	Нет
Учет регистра слов	Нет, кроме Слов с большой буквы	Нет, кроме Слов С Большой Буквы , следующих подряд	

\	Яндекс	Rambler	Google
Учет расстояний между словами	"Фразы в кавычках", Слово1 /N Слово2 - расстояние между словами не больше N слов, Слово1 /(-N M) Слово2 - расстояние между словами от -N слов слева и M слов справа, Слово1 &&/(-N M) Слово2 - расстояние между словами от -N предложений слева и M предложений справа	"Фразы в кавычках", (N, Слово1 Слово2 ...) - расстояние между словами не более N слов (В противном случае расстояние не более 40 слов между словами запроса; Без кавычек и скобок предлоги, частицы и артикли игнорируются)	"Фразы в кавычках"
Расширенные операторы	\$title \$anchor #keywords #abstract #image #hint #url #link *	-	cache: link: related: info: stocks: site: all: intitle: intitle: all: nurl: inurl: admission site: Language: Date: Occurrences: Domains: SafeSearch:
Уточнение поиска	Есть	Есть	Нет
Поиск в расширенном режиме	Словарный фильтр, Дата, Сайт/вершина, Ссылка, Изображение, Специальные объекты, Язык, Формат выдачи	Поиск по тексту, Искать слова запроса, Расстояние между словами запроса, Исключить документы содержащие следующие слова, Язык документа, Искать документы только на следующих сайтах, Вывод результатов поиска, Сортировать, Выдавать, Форма вывода, Связанные запросы	Найти результаты (со всеми словами, с точной фразой, с любым из слов, без слов), Язык, Формат файла, Дата, Упоминание, Домен, Поиск по странице (Похожие, Ссылки)
Объекты поиска	WWW, Каталог, Новости, Маркет, Карты, Словари, Блоги, Картинки, RTF, PDF	WWW, FTP, Новости, Картинки, Покупки, Price.ru, Топ100, Словари	WWW, Картинки, Карты, Новости, Группы, PDF
Дополнительные возможности			Перевод страницы, Похожие страницы

Описание языка запросов конкретных поисковых систем (ссылки)

8.5. Обзор поисковых систем

Google <http://www.google.com>

Начало этой поисковой службе с непонятным названием положили студенты Стэндфордского университета: бывший москвич Сергей Брин и Ларри Пейдж. Слово «Google» является производным от «термина» googol, придуманного племянником известного американского математика Эдварда Каснера, и означает число, записываемое как единица со ста нулями.

Академичность происхождения определила академичность подхода. Компаньоны взяли на вооружение общеизвестную систему «оценки ценности» статей, принятую в мировом научном сообществе.

Рейтинг статьи есть производная от количества сделанных цитат и ссылок на нее в других научных публикациях.

Google высчитывает релевантность документа, попавшего в результаты поиска, в соответствии с количеством

вом ссылающихся на него других Web-страниц. «Старинные» бумажные принципы оказались действенными и в Интернете.

Сбором данных в системе занимаются несколько независимых роботов, получающих задание от сервера, коллекционирующего ссылки. Найденные документы архивируются и индексируются по словам, документам и ссылкам.

Составляя запрос, следует помнить, что Google:

- 1) автоматически вставляет между всеми словами запроса оператор AND и не поддерживает оператор OR, а также возвращает только те страницы, которые содержат абсолютно все термины запроса;
- 2) позволяет исключить слова-«паразиты» с помощью символа "-" и не работает с шаблонами (?, * и т.д.);
- 3) рассматривает слова, заключенные в двойные парные кавычки, как указание искать только полные совпадения фраз и автоматически удаляет отдельно стоящие цифры и буквы, слова com, http и тому подобные «бессмысленные» термины, но сделает исключение для вас, если обнаружит значок «+» перед ними;
- 4) предлагает особый вид услуг - выдачу всех ссылок на заданную страницу (link: <url>).

На первой же странице результатов поиска Google можно получить сведения об общем количестве найденных документов и времени, затраченном на обработку запроса. Далее идут собственно ссылки на них, сгруппированные по сайтам.

Функция GoogleScout поможет отыскать все похожие документы (аналогично многим российским поисковым машинам).

Компания Google не ограничилась только поисковым сервисом и занимается созданием собственного Интернет-каталога под названием GoogleDirectory. На сегодняшний день GoogleDirectory насчитывает описания более полутора миллионов специально отобранных узлов. Информация, почерпнутая из каталога, всегда дополняет результаты традиционного поиска в Web, так что дважды искать вам не придется.

Яндекс. <http://www.yandex.ru> и <http://www.ya.ru>

Яндекс — поисковая машина, способная по вашему запросу найти наиболее подходящие веб-страницы в русской части Интернета. Яндекс ежедневно просматривает сотни тысяч веб-страниц в поисках изменений или новых ссылок.

Яндексу можно задать вопрос так же, как бы вы его задали библиотекарю или всезнайке-эрудиту. Например, «где раки зимуют», «ярчайшая звезда северного полушария» или «как выбрать компьютер».

Советы по поиску в Яндексе:

- 1) Проверяйте орфографию.
- 2) Используйте синонимы. Задайте для поиска три-четыре слова-синонима сразу, перечислив их через вертикальную черту (|). Тогда будут найдены страницы, где встречается хотя бы одно из них. Например, вместо «фотографии» попробуйте «фотографии | фото | фотоснимки».
- 3) Ищите больше, чем по одному слову. Слово «психология» или «продукты» дадут при поиске по одиночке большое число бессмысленных ссылок. Добавьте одно или два ключевых слова, связанных с искомой темой. Например, «психология Юнга» или «продажа и покупка продовольствия».
- 4) Не пишите большими буквами. Начиная слово с большой буквы, вы не найдете слов, написанных с маленькой буквы, если это слово не первое в предложении. Поэтому не набирайте обычные слова с Большой Буквы, даже если с них начинается ваш вопрос Яндексу. Заглавные буквы в запросе рекомендуется использовать только в именах собственных.
- 5) Найти похожие документы. Если один из найденных документов ближе к искомой теме, чем остальные, нажмите на ссылку «найти похожие документы». Ссылка расположена под краткими описаниями найденных документов. Яндекс проанализирует страницу и найдет документы, похожие на тот, что вы указали.
- 6) Используйте знаки «+» и «-». Чтобы исключить документы, где встречается определенное слово, поставьте перед ним знак минуса. И наоборот, чтобы определенное слово обязательно присутствовало в документе, поставьте перед ним плюс. Обратите внимание, что между словом и знаком плюс-минус не должно быть пробела. Плюс стоит использовать в том случае, когда нужно найти так называемые стоп-слова (наиболее частотные слова русского языка, в основном это местоимения, предлоги, частицы). Чтобы найти цитату из Гамлета, надо задать запрос «+быть или +не быть».
- 7) Попробуйте использовать язык запросов. С помощью специальных знаков вы сможете сделать запрос более точным. Например, укажите, каких слов не должно быть в документе, или что два слова должны идти подряд, а не просто оба встречаться в документе.
- 8) Поиск картинок и фотографий. Яндекс умеет искать не только в тексте документа, но и отыскивать картинки по названию файла или подписи. Для этого на первой странице yandex.ru нажмите ссылку «расширенный поиск».

9) Независимо от того, в какой форме вы употребили слово в запросе, поиск учитывает все его формы по правилам русского языка. Например, если задан запрос 'идти', то в результате поиска будут найдены ссылки на документы, содержащие слова 'идти', 'идет', 'шел', 'шла' и т.д.

Чтобы искать без морфологии надо поставить «!» перед словом. Например, запрос !Иванов найдет только страницы с упоминанием этой фамилии, а не города «Иваново».

При этом поиск не ограничен лишь словами или фразами. Яндекс отыщет всех, кто сослался на вашу страницу, файлы с нужной картинкой, последние новости или товары в электронных магазинах.

Результаты поиска.

После того, как вы задали запрос, Яндекс выведет список ссылок на Документы, наиболее точно ему соответствующие. На этой странице результатов вы обнаружите некоторые специальные обозначения и ссылки, предназначенные облегчить просмотр и сортировку найденных страниц.

8.6. Вопросно-ответная система

Вопросно-ответная система (англ. Question-answering system) — это особый тип информационных систем, являющиеся гибридом поисковых, справочных и интеллектуальных систем (часто они рассматриваются как интеллектуальные поисковые системы).

QA-система должна быть способна принимать вопросы на естественном языке, то есть это система с естественно-языковым интерфейсом. Информация предоставляется на основе документов из сети Интернет или из локального хранилища. Современные разработки QA-систем позволяют обрабатывать множество вариантов запросов фактов, списков, определений, вопросов типа Как, Почему, гипотетических, сложных и межязыковых.

Узкоспециализированные QA-системы работают в конкретных областях (например, медицина или обслуживание автомобилей). Построение таких систем — сравнительно легкая задача.

Общие QA-системы работают с информацией по всем областям знаний, таким образом появляется возможность вести поиск в смежных областях.

Архитектура

Первые QA-системы были разработаны в 1960х годах и являлись естественно-языковыми оболочками для экспертных систем, ориентированных на конкретные области. Современные системы предназначены для поиска ответов на вопросы в предоставляемых документах с использованием технологий обработки естественных языков (NLP).

Современные QA-системы обычно включают особый модуль — классификатор вопросов, который определяет тип вопроса и, соответственно, ожидаемого ответа. После этого анализа система постепенно применяет к предоставленным документам все более сложные и тонкие методы NLP, отбрасывая ненужную информацию. Самый грубый метод — поиск в документах — предполагает использование системы поиска информации для отбора частей текста, потенциально содержащих ответ. Затем фильтр выделяет фразы, похожие на ожидаемый ответ (например, на вопрос «Кто ...» фильтр вернет кусочки текста, содержащие имена людей). И, наконец, модуль выделения ответов найдет среди этих фраз правильный ответ.

Схема работы

Производительность вопросно-ответной системы зависит от качества текстовой базы — если в ней нет ответов на вопросы, QA-система мало что сможет найти. Чем больше база — тем лучше, но только если она содержит нужную информацию. Большие хранилища (такие как Интернет) содержат много избыточной информации. Это ведёт к двум положительным моментам:

Так как информация представлена в разных формах, QA-системе быстрее найдет подходящий ответ. Не придется прибегать к сложным методам обработки текстов.

Правильная информация чаще повторяется, поэтому ошибки в документах отсеиваются.

Поверхностный поиск

Наиболее распространенный метод поиска — по ключевым словам. Найденные таким способом фразы фильтруются в соответствии с типом вопроса, а затем ранжируются по синтаксическим признакам, например, по порядку слов.

Расширенный поиск

В 2002 году группа исследователей написала план исследований в области вопросно-ответных систем. Предлагалось рассмотреть следующие вопросы.

1. Типы вопросов

Разные вопросы требуют разных методов поиска ответов. Поэтому нужно составить или улучшить методические списки типов возможных вопросов.

2. Обработка вопросов

Одну и ту же информацию можно запросить разными способами. Требуется создать эффективные методы понимания и обработки семантики (смысла) предложения. Важно, чтобы программа распознавала эквивалентные по смыслу вопросы, независимо от используемых стиля, слов, синтаксических взаимосвязей и идиом. Хотелось бы, чтобы QA-система разделяла сложные вопросы на несколько простых, и правильно трактовала контекстно-зависимые фразы, возможно, уточняя их у пользователя в процессе диалога.

3. Контекстные вопросы

Вопросы задаются в определенном контексте. Контекст может уточнить запрос, устранить двусмысленность или следить за ходом мыслей пользователя по серии вопросов.

4. Источники знаний для QA-системы

Перед тем как отвечать на вопрос, неплохо было бы осведомиться о доступных базах текстов. Какие бы способы обработки текстов не применялись, мы не найдем правильного ответа, если его нет в базах.

5. Выделение ответов

Правильное выполнение этой процедуры зависит от сложности вопроса, его типа, контекста, качества доступных текстов, метода поиска и др. — огромного числа факторов. Поэтому подходить к изучению методов обработки текста нужно со всей осторожностью, и эта проблема заслуживает особого внимания.

6. Формулировка ответа

Ответ должен быть как можно более естественным. В некоторых случаях достаточно и простого выделения его из текста. К примеру, если требуется наименование (имя человека, название прибора, болезни), величина (денежный курс, длина, размер) или дата («Когда родился Иван Грозный?») — прямого ответа достаточно. Но иногда приходится иметь дело со сложными запросами, и здесь нужны особые алгоритмы слияния ответов из разных документов.

7. Ответы на вопросы в реальном времени

Нужно сделать систему, которая бы находила ответы в хранилищах за несколько секунд, независимо от сложности и двусмысленности вопроса, размера и обширности документной базы.

8. Многоязыковые запросы

Разработка систем для работы и поиска на других языках (в том числе автоматический перевод).

9. Интерактивность

Зачастую информация, предлагаемая QA-системой в качестве ответа, неполна. Возможно, система неправильно определила тип вопроса или неправильно «поняла» его. В этом случае пользователь может захотеть не только переформулировать свой запрос, но и «объясниться» с программой с помощью диалога.

10. Механизм рассуждений (вывода)

Некоторые пользователи хотели бы получить ответ, выходящий за рамки доступных текстов. Для этого в QA-систему нужно добавить знания, общие для большинства областей (см. Общие онтологии в информатике), а также средства автоматического вывода новых знаний.

11. Профили пользователей QA-систем

Сведения о пользователе, такие как область интересов, манера его речи и рассуждения, подразумеваемые по умолчанию факты, могли бы существенно увеличить производительность системы.

8.7. Коллаборативная, или Совместная фильтрация

Современного человека окружает огромное информационное пространство, в котором довольно легко заблудиться, но еще сложнее в нем найти то, что действительно заинтересует именно вас. Другая проблема большого объема информации в том, что вам хочется усвоить как можно больше полезных сведений, соответствующих вашим предпочтениям. Искать и перебирать книги, фильмы, спектакли — это довольно длительный процесс. Он отнимает драгоценное время и деньги: покупая книгу или билет на зрелище, никто не может гарантировать, что он получит удовольствия от процесса ознакомления с этим продуктом.

Конечно, вы не одни в этом мире. У вас есть знакомые и друзья, которые вам могут порекомендовать что-либо интересное, можно также ознакомиться с рецензиями различных критиков. Однако у друзей, хоть они и ваши друзья, вкус может быть совершенно отличным от вашего, а критики — тоже люди и не могут охватить весь объем новинок. Рейтинги и модные авторы — это настолько временное явление, что года через три-четыре можно почувствовать жгучий стыд за то, что вы когда-то это читали или превозносили того или иного автора, актера, книгу, фильм.

Что же делать, как ориентироваться в потоке информации? Вам поможет коллаборативная фильтрация.

Коллаборативная, или Совместная фильтрация (collaborative filtration) — метод, дающий автоматические прогнозы (фильтрацию) относительно интересов пользователя по собранной информации о вкусах множества пользователей (сотрудничающих между собой).

В мире коллаборативные системы создают с 1994 года, имеется обширная литература, проводятся конференции. Сотни систем такого рода с переменным успехом работают для разных типов контента – от книг и фильмов до новостей и вин.

По общему признанию, самыми продвинутыми на сегодня являются:

- система крупнейшего мирового книжного магазина Amazon.com, основанная на item-based-фильтрации («Что покупали те, кто покупал Это»),
- система Cinematch компании NetFlix.com (прокат DVD) – на основе user-based фильтрации.

К сожалению, пока еще не существует системы, которая бы, например, на основе оцененных книг и музыкальных произведений могла бы вам порекомендовать, какой на досуге посмотреть фильм.

А так принцип этого метода довольно прост:

- создается база однотипных объектов (книги, музыка, фильмы, веб-сайты и т.п.);
- задается шкала оценок (пятибалльная, десятибалльная, стобалльная);
- множество пользователей проставляет оценки данным объектам по своему усмотрению;
- каждый пользователь, проставляя свои оценки, задает тем самым системе свой индивидуальный пользовательский профиль;
- на основе оценок объектов, проставленных всеми пользователями, и соответствия оценок конкретного пользователя оценкам других профилей клиент получает список рекомендуемых к ознакомлению объектов с приблизительной оценкой каждого из них.

Это субъект-субъектная система. Вы практически получаете информацию от своих единомышленников и даете им возможность познакомиться с тем, что является для вас уже пройденным этапом, а им только предстоит. Система экономит ваше время на поиски, гарантированно предлагая вам заведомо интересную книгу или фильм.

Преимущество этих сообществ в том, что вы можете не только найти интересные для себя культурные объекты, но и пообщаться с людьми схожих с вашими вкусов, найти среди них поддержку, посоветоваться, не сходя с места. Это особенно актуально в сегодняшнем мире, когда на привычные кружки по интересам уже не остается сил и времени.

Важно! Прибегая к помощи подобных сервисов, нужно помнить одну очень простую вещь: чем больше вы поставите оценок, тем лучше и точнее получите рекомендацию, тем приятнее, полезнее и интереснее проведете время.

И, конечно же, не обязательно слепо следовать рекомендациям. Вы можете читать, смотреть и слушать все, что вашей душе угодно.

8.8. Электронная библиотека

Электронная библиотека — упорядоченная коллекция разнородных электронных документов, снабженных средствами навигации и поиска.

Может быть веб-сайтом, где постепенно накапливаются различные тексты (чаще литературные, но также и любые другие, вплоть до компьютерных программ) и медиа-файлы, каждый из которых самодостаточен и в любой момент может быть востребован читателем.

Электронные библиотеки могут быть универсальными, стремящимися к наиболее широкому выбору материала (как Библиотека Максима Мошкова), и более специализированными, как Фундаментальная электронная библиотека или проект Сетевая Словесность, нацеленный на собирание авторов и типов текста, наиболее ярко заявляющих о себе именно в Интернете.

Особое место в ряду электронных библиотек занимают интернет-библиотеки научно-образовательной тематики (например Электронная библиотека IqLib), в которых собраны издания, необходимые для осуществления образовательного процесса.

Форматы выкладываемых произведений

Форматы размещаемых в электронных библиотеках произведений можно разделить на две категории — форматы, предназначенные для чтения текста он-лайн и форматы, предназначенные для скачивания на компьютер читателя.

Самый популярный **формат первой категории** — HTML, однако библиотека Мошкова, например, выкладывает тексты в формате TXT. Связано это с традицией: во время появления этой библиотеки (1994 год) скорости работы в интернет были весьма низки, и оправданным было использование самого легкого формата. Также по философии библиотеки каждое произведение должно было целиком уместиться на стандартном носителе

информации (каковым тогда являлась дискета).

Форматы для скачивания — заархивированный TXT; RTF и собственный двоичный формат файлов Microsoft Word 97—2000; Mobipocket.PRC (формат для чтения книг на кпк и телефонах).

В последнее время популярность набирает формат FictionBook, созданный специально для хранения литературных произведений.

Первая электронная библиотека — Проект «Гутенберг» — требует в файлах для скачивания использовать текстовый формат в архивах ZIP для того, чтобы тексты можно было читать практически на любых устройствах, и чтобы они не были потеряны в случае исчезновения программ для обработки используемых форматов.

Книги, изобилующие математическими формулами и сложными схемами, после сканирования переводить в текстовый формат намного сложнее, поэтому часто их хранят в графическом формате, обычно DjVu и PDF. Тогда как PDF при таком применении представляет собой просто объединенный в один файл набор изображений TIFF, DjVu использует специальный алгоритм, позволяющий получать в несколько раз меньшие файлы даже при сжатии без потерь.

Наоборот, если существует электронный оригинал книги — с редактируемым текстом и векторными изображениями, то именно PDF будет лучшим выбором.

Электронные библиотеки, нарушающие авторские права

В рунете сложилась уникальная ситуация, связанная с авторскими правами на литературные произведения. Многие электронные библиотеки публикуют литературные произведения без предварительного согласия авторов, и хотя в некоторых из них по первому требованию автора его тексты снимаются с сайта, юридически они нарушают законы об авторском праве, однако государство предпринимает мало усилий (если вообще предпринимает) для борьбы с такими нарушениями. Пользователи рунета также, как правило, на стороне «нарушителей прав издателей», и неодобрительно относятся к попыткам авторов или издателей судиться с электронными библиотеками.

Аргументы сторонников этого похожи на аргументы незаконных пользователей собственного ПО:

Электронные библиотеки служат рекламой для авторов так как дают возможность прочитать их книги той части аудитории, которая не может позволить себе их покупать.

Электронные библиотеки публикуют произведения, с которыми нельзя ознакомиться другим способом (например, книга давно не переиздавалась или издавалась малым тиражом).

Существующая в России система авторского права далека от совершенства, что мешает авторам регулировать степень доступности своих произведений.

Многие люди считают что электронные книги удобнее бумажных, но не имеют возможности легально получить их.

Списки библиотек и поисковые системы

- [Электронные библиотеки в DMOZ](#)
- [Закрома Родины или Где в И-нете Пылятся Электронные Книги](#) — обширный список электронных библиотек.
- [Библио](#) — поисковая система в электронных библиотеках
- [DjVu библиотеки](#) — научные и технические библиотеки.
- Универсальные**
- [Проект «Гутенберг»](#) — первая в мире электронная библиотека. Содержит более 20000 текстов, аудиокниг, [нотные записи](#), MIDI и иногда видеороликов, в основном, находящихся в [общественном достоянии](#) в США.
- [LibriVox](#) — проект по созданию [аудиокниг](#) в общественном достоянии, на основе текстов, находящихся в ОД в США — в основном, из проекта «Гутенберг».
- [Библиотека Мошкова](#) — старейшая, наиболее известная и одна из крупнейших русскоязычных сетевых библиотек. Содержит литературные произведения всех жанров, а также нехудожественную литературу и музыку.
- [Библиотека ImWerden](#)

8.9. Проблемы в работе поисковых систем

Улучшение работы поисковых систем — это одна из приоритетных задач сегодняшнего Интернета. Основные проблемы в работе поисковых систем связаны с так называемой Глубокой паутиной.

8.9.1. Глубокая паутина

Глубокая паутина (также известна как невидимая паутина) — множество веб-страниц Всемирной паутины, неиндексируемых поисковыми системами.

Термин произошел от соотв. англ. *invisible web*. Наиболее значительной частью глубокой паутины является глубинный веб (от англ. *deep web*, *hidden web*), состоящий из веб-страниц, динамически-генерируемых по запросам к онлайн базам данных. Не следует смешивать понятие глубокая паутина с понятием тёмная паутина (от англ. *dark web*), под которым имеются в виду сетевые сегменты, вообще не подключённые к сети Интернет.

Суть проблемы

В глубокой паутине находятся веб-страницы, несвязанные с другими гиперссылками — например, страницы, динамически создаваемые по запросам к базам данных. В глубокой паутине также находятся сайты, доступ к которым открыт только для зарегистрированных пользователей.

Поисковые системы используют специальные роботы (англ. *web crawler*), которые переходят по гиперссылкам и индексируют содержимое веб-страниц, на которых они оказываются. В то же время, информация из баз данных, доступная пользователям через поисковые веб-формы (но не по гиперссылкам), остается недоступной для робота, неспособного в режиме реального времени правильно заполнить форму значениями (другими словами, сформировать запрос к базе данных).

Таким образом, значительная часть Всемирной паутины оказывается скрыта от поисковых роботов. Используя аналогию, информация, будучи недоступной для поисковых систем, находится «на глубине» (от англ. *deep*).

Масштаб

В 2000 году поисковая компания «BrightPlanet» провела исследование, которое показало, что в глубокой паутине находится в 500 раз больше документов, чем доступно через поисковые системы. Конечно, эти цифры могут быть не совсем точными. Например, существует проблема с различием разных представлений одного и того же материала в базах данных.

Ключи к решению

В 2005 году компания «Yahoo!» сделала серьезный шаг к решению этой проблемы. Компания выпустила поисковый движок «Yahoo! Subscriptions», который производит поиск по сайтам (пока немногочисленным), доступ к которым открыт только зарегистрированным участникам этих сайтов.

Это, однако, полностью не решило имеющейся проблемы. Эксперты поисковых систем по-прежнему пытаются найти технические возможности для индексации содержимого баз данных и доступа к закрытым веб-сайтам.

Информационные источники

1. Саак А.Э., Пахомов Е.В., Тюшняков В.Н. **Информационные технологии управления**: Учебник для вузов + CD. - СПб.: Питер, 2008