

Интеллектуальные информационные системы

Лекция 3

Инструментальные средства создания гипертекстовых систем

Благодаря широкому использованию ГТ в ИС практически любой инструментарий разработки ИС включает функции для построения ГТ. В частности, данные функции реализуются в средствах разработки электронной документации (например, Adobe Acrobat), авторских системах, редакторах презентаций, издательских системах, редакторах web-страниц и др.

Существует также специализированный инструментарий:

Microsoft Windows Help (WinHelp) и HTML Help

- стандартные технологии построения и работы с гипертекстовыми справочниками для платформы Windows. Они позволяют формировать самые разнообразные ГТ:
 - электронные руководства,
 - справочники,
 - энциклопедии,
 - пособия и др.

Однако главное назначение данных технологий — реализация контекстно-зависимых гипертекстовых справочников по программным продуктам. Такие справочники являются неотъемлемым компонентом прикладных программных систем. По умолчанию они вызываются клавишей F1 или через меню «Справка». Информация, отображаемая в окне справочника после его вызова, зависит от текущего режима работы приложения, с которым он связан. Поэтому подобные справочники называются контекстно-зависимыми

Создание гипертекстового справочника по программному продукту состоит из шести основных этапов.

1. Определение структуры справочника и его разделов.

Этот этап является наиболее сложным и трудно формализуемым. В рамках него специфицируются:

- назначение продукта, для которого создается справочник;
- категории пользователей продукта;
- рыночный сектор, на который ориентирован продукт;
- функции и характеристики продукта, представляемые в справочнике;
- основные разделы справочника и их примерное содержание;
- соглашения, фиксирующие стиль, дизайн и оформление справочника.

2. Подготовка текста и графических иллюстраций справочника.

Определение гипертекстовых ссылок.
Формирование файлов тем (ИСС) и графических файлов, включая задание контактных областей для гиперграфики.

3. Создание файла проекта справочника.

4. Компиляция исходных файлов тем, графических файлов и файла проекта с формированием файла справочника.
5. Программная реализация модуля приложения, обеспечивающего доступ к справочнику.
6. Тестирование и отладка справочника.

Гипертекст в формате WinHelp реализуется в виде файла с расширением HLP (help-файла). Представление и взаимодействие со справочником обеспечивает программа WINHELP.EXE, входящая в состав Windows. HLP-файл формируется на основе файлов с текстом в формате RTF с помощью специального компилятора. Для вызова справочника из приложения служит функция Windows API WinHelp().

Гипертекст в формате HTML Help реализуется в виде файла с расширением CHM. Представление и взаимодействие со справочником обеспечивают программные компоненты браузера Internet Explorer (начиная с версии 4.0). Для вызова справочника из приложения служит функция HTML Help API HtmlHelp().

К достоинствам HTML Help

относятся:

- мощные средства языка HTML, включая каскадные таблицы стилей;
- возможности использования компонентов ActiveX и скриптов;
- тесная интеграция с технологиями Internet;
- возможность создания составных гипертекстовых справочников, объединяемых во время выполнения.
- Информация в СНМ-файле хранится в сжатом виде. Степень компрессии составляет примерно 8:1.

Гипертекст в формате HTML Help может быть разработан с помощью различных инструментальных средств. Наиболее популярными из них являются HTML Help Workshop фирмы Microsoft и KeyTools фирмы KeyWorks Software. Система Anet Help Tool российской фирмы Anet Soft позволяет создавать ГТ в формате как HTML Help, так и WinHelp.

Инструментальная среда HyperRef

Предназначена для построения электронных гипертекстовых изданий большого объема. Разработана в МЭИ (ТУ).

HyperRef поддерживает следующие типы информационных объектов:

- текстовые экранные страницы,
- графические изображения,
- исполняемые модули.

Инструментальная среда HyperRef

Объекты объединяются как в линейные последовательности, метафорой которых является глава или раздел книги, так и в гипертекстовую сеть. В визуальных объектах могут быть определены интерактивные элементы, используемые для организации гиперссылок.

HyperRef поддерживает типизацию гиперссылок и содержит средства навигации по ГТ с учетом ограничений, обусловленных типами ссылок.

В состав HyperRef входят:

- диалоговый инструментарий автора (конструктор);
- пользовательская программа для работы с ГТ (исполнитель);
- набор утилит, позволяющих осуществлять поточный ввод информации, контролировать и восстанавливать целостность электронных гипертекстовых документов и т. д.

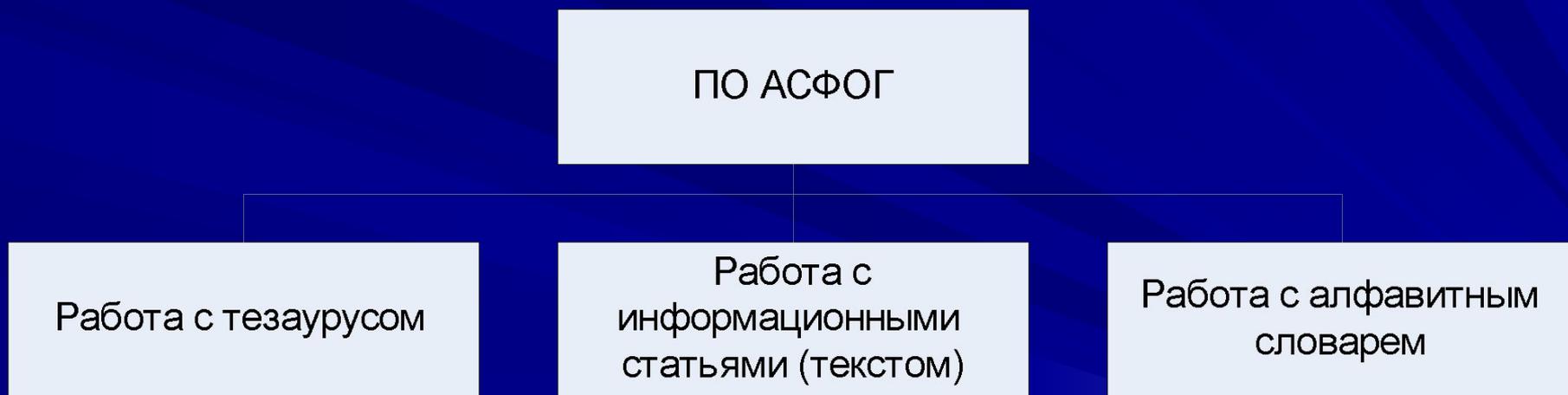
В HyperRef предусмотрены средства, присущие фактографическим и полнотекстовым БД: словари ключевых слов, оглавления, средства выполнения сложных запросов и автоматической индексации текстов.

Автоматизированная система формирования и обработки гипертекстов (АСФОГ)

создана в МЭСИ, предназначена для моделирования экономических объектов и процессов на основе представления информационного фонда ПрО в виде ГТ.

АСФОГ целесообразно использовать для моделирования слабоструктурированных ПрО, когда поиск текстовой информации в традиционных линейных и иерархических структурах неэффективен из-за их неадекватности реальной сетевой структуре информационных объектов, представляющих эти ПрО.

Программное обеспечение АСФОГ реализовано в трех подсистемах



Подсистема работы с тезаурусом

выполняет следующие функции:

- поиск в тезаурусе (поиск по связям с учетом их типов, контекстный поиск по связям);
- поддержка ускоренного просмотра;
- формирование отчетов;
- поддержка формирования и корректировки тезауруса.

Подсистема работы с информационными статьями

- создание ИСС с помощью текстового редактора типа Word;
- коррекция ИСС;
- доступ к ИСС;
- формирование и печать отчетов по ИСС;
- импорт и экспорт файлов, содержащих ИСС.

Подсистема работы с алфавитным словарем решает следующие задачи:

- алфавитная сортировка (лексико-графическое упорядочение) заголовков ИСС;
- контекстный поиск ИСС по заголовку;
- поддержка ускоренного просмотра словаря;
- печать информации из словаря.

Гипертекстовые информационно-поисковые системы

Гипертекстовая информационная технология используется при организации больших массивов текстовых документов и реализации методов поиска информации в них.

Информационный поиск — совокупность операций, методов и процедур, направленных на отбор данных, хранящихся в ИС и соответствующих заданным условиям.

Информационно-поисковые системы (ИПС) подразделяются на три класса:

- документальные;
- фактографические;
- гипертекстовые (ГИПС).

Документальные ИПС

Документальные ИПС хранят и выдают сведения о документах, основное содержимое которых представлено в виде связанного текста на естественном языке (ЕЯ). Признаки документа, отражающие его содержание в ИПС, называют поисковым образом, а признаки запроса к ИПС — поисковым предписанием.

Процедура перевода документа и запроса в форму представления, принятую в ИПС, называется индексированием. При сопоставлении поискового образа и поискового предписания используется тот или иной критерий смыслового соответствия (релевантности).

Первые ИПС были предназначены для поиска книг в библиотеках и получили название библиографических. Позже их стали применять и для поиска документов в больших хранилищах и стали называть документальными

Основным объектом информационного фонда документальной ИПС является аннотация (реферат) и библиографическое описание документа (книги, события, предмета). Реферат (аннотация) выражается на ЕЯ и отражает основные характеристики документа, представляющие интерес для пользователей. Предполагается, что в подобном описании можно выделить ряд слов и словосочетаний, число которых значительно меньше общего числа слов в описании. В то же время выделенная информация достаточно точно характеризует описание. Такие слова и словосочетания называются ключевыми словами или дескрипторами.

Запрос к документальной ИПС формулируется в виде перечня дескрипторов, которые по мнению пользователя характеризуют искомый документ.

При вводе в ИПС нового объекта (реферата) его дескрипторы автоматически включаются в словарь дескрипторов. Каждому дескриптору присваивается номер, называемый индексом дескриптора. Совокупность индексов, соответствующих полному набору дескрипторов реферата, составляет его поисковый образ. Новый поисковый образ снабжается уникальным идентификатором (регистрируется) и включается в массив поисковых образов. Тем же идентификатором помечается новый реферат, заносимый в массив рефератов.

Организация поиска в дескрипторной ИПС

Запрос, сформулированный на ЕЯ, подвергается анализу, в рамках которого в нем выделяются дескрипторы, входящие в словарь дескрипторов. Их совокупность образует поисковое предписание, соответствующее запросу. Оно сопоставляется с поисковыми образами, в результате чего определяется их релевантность. Если поисковый образ и предписание релевантны, то из поискового образа извлекается идентификатор реферата, выдаваемого пользователю. Ответом на запрос является множество рефератов, соответствующих отобранному в процессе поиска идентификаторам.

В целях ускорения поиска для каждого дескриптора в словаре дескрипторов указывается список идентификаторов рефератов, в которых он встречается. Такая информационная структура ИПС называется индексом.

С помощью дескрипторов можно лишь приблизительно отразить смысл документов. Это же относится к переводу запросов в поисковые предписания. Документальная ИПС может выдать рефераты, не относящиеся к поисковому запросу, или не найти рефераты, которые соответствуют ему.

Документальный поиск относится к числу сложных информационных процессов, поскольку он связан с проблемой оценивания смыслового соответствия документа и запроса. Из-за субъективности и неоднозначности подобного оценивания этот вид поиска в принципе не может быть исчерпывающе точным и полным, в нем всегда будет присутствовать элемент нечеткости.

Развитием поиска по дескрипторам является полнотекстовый поиск, реализуемый, например, в поисковых машинах Internet. В системах, использующих данный вид поиска, индекс формируется на основе всех слов и словосочетаний, содержащихся в документах, за исключением служебных неинформативных слов (союзов, предлогов, местоимений и т. п.). При индексировании с помощью словарей и средств морфологического анализа слова приводятся к базовой грамматической форме (именительный падеж, единственное число и т. д.).

Фактографические ИПС

В фактографических ИПС хранятся не документы, а собственно сведения (факты) об объектах ПрО. Подобные ИПС реализуются, в частности, на основе реляционных БД. С точки зрения обеспечения релевантности результатов поиска (выборки данных) запросу фактографический поиск в отличие от документального является ТОЧНЫМ И ПОЛНЫМ.

Гипертекстовые ИПС

В гипертекстовых ИПС кроме содержимого документов отражается их семантическая структура. Поэтому по глубине формализации ГИПС занимают промежуточное положение между документальными и фактографическими ИПС.

Поиск по метаданным

Одно из направлений развития технологии документальных ИПС связано со структуризацией и унификацией сведений о документах. Такие сведения по отношению к исходным документам играют роль метаданных. Примером метаданных служит библиографическое описание, содержащее информацию об авторах документа, дате его создания, объеме, форме представления и т. д. Ключевые слова также относят к метаданным.

Поиск по метаданным сближает технологии документальных и фактографических ИПС. С одной стороны, метаданные представляют документы. С другой стороны, некоторые элементы метаданных допускают четкое определение релевантности запроса и записи в БД (экземпляра метаданных, ассоциируемых с конкретным документом), что характерно для фактографических ИПС.

В настоящее время хранилища метаданных обычно реализуются на основе реляционных и XML-ориентированных БД и используют механизмы поиска, воплощаемые в соответствующих СУБД.

Классификация и характеристики методов информационного поиска



Введем следующие обозначения:

D – множество документов в информационном хранилище

$d_i \in D$ – i – тый документ

$D_j \in D$ – подмножество документов

В данном контексте под документом будем понимать как собственно текстовый или гипертекстовый документ, так и отдельную запись в БД.

Зададим на D оценку смысловой близости пары документов $r(d_i, d_j) \geq 0$. При $r=0$ документы d_i и d_j эквивалентны по смыслу. Для семантически несопоставимых документов r не определена.

Введем оценки ряда важных свойств документов: $S=(S_1, S_2, \dots, S_k)$, $k>0$. Пусть оценка каждого свойства S_i выражается действительным числом, принадлежащим некоторому интервалу. Для определенности примем, что чем больше значение S_i , тем важнее для пользователя документ.

Поисковый запрос может рассматриваться как виртуальный документ z . В идеальном случае ($r(z, d_j) = 0$) ему точно соответствует документ d_j .

Используя введенные обозначения, определим следующие виды поиска:

1. Найти $(D_j \subseteq D) \mid r(z, d_i \in D_j) \rightarrow \min$

Если $D_j = \emptyset$, то в D нет документов, релевантных запросу. При $|D_j|=1$ есть единственный подходящий документ. Если $|D_j|>1$, то таких документов несколько

2. Найти $(D_j \subseteq D) \mid r(z, d_i \in D_j) \leq \Delta$ — оценка наибольшего допустимого расхождения смыслов запроса и искомым документов.

3. Найти $(D_j \subseteq D) | S_j(d_i \in D_j)$ Результатом поиска служит подмножество документов, которым приписана наибольшая оценка важности j -го свойства. Обобщением этого варианта является векторный поиск, учитывающий оценки нескольких свойств.

4. Комбинированный поиск: найти

$(D_j \subset D) | r(z, d_i \in D_j) \leq \Delta \& S_j(d_i \in D_j) \rightarrow \max$
Интеллектуальные возможности ИГТС в части функций информационного поиска обусловлены способами задания и вычисления r и S .

Эффективность информационного поиска документов, обеспечиваемая ИПС, оценивается по информационной полноте и информационному шуму. Названные показатели выражаются коэффициентами полноты k_i и шума k_o соответственно. Коэффициенты k_i и k_o принимают значения в интервале от 0 до 1. В некоторых источниках эти коэффициенты выражают в процентах.

Пусть ИПС предъявлен i -й запрос. Информационно-поисковая система содержит множество документов D_i релевантных этому запросу. В результате поиска получено множество D_i^o , Возможны следующие варианты:

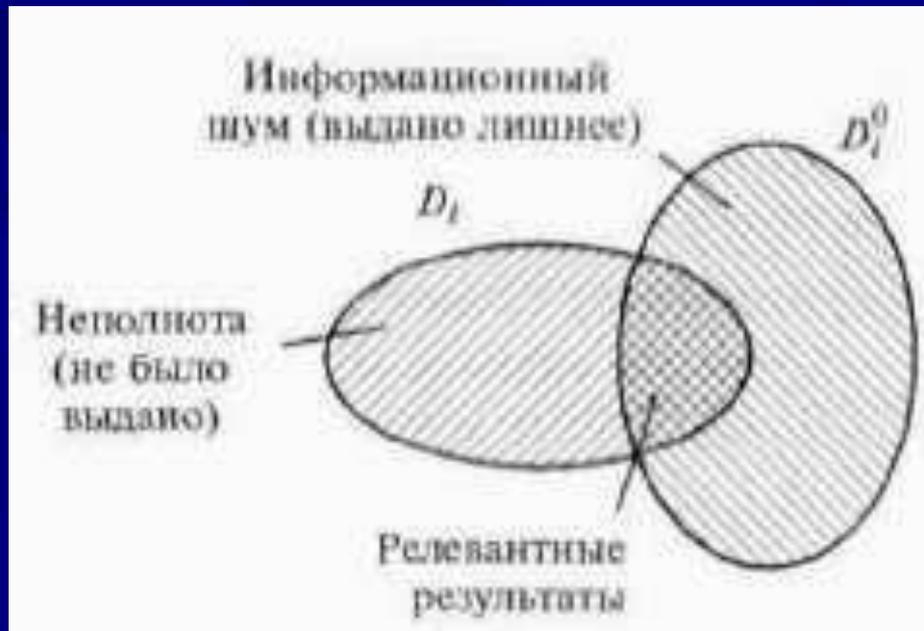
1. $D_i^0 = D_i$. Идеальный вариант: полнота максимальна ($k_r = 1$) а шум нулевой ($k_g = 0$).
2. $D_i^0 \subset D_i$. Имеет место неполнота ($0 \leq k_r < 1$), а шум отсутствует ($k_g = 0$).
3. $D_i^0 \supset D_i$. Неполнота исключается ($k_r = 1$), но есть шум ($0 \leq k_g < 1$).
4. $D_i^0 \cap D_i = \emptyset \ \& \ D_i^0 \neq \emptyset \ \& \ D_i \neq \emptyset$. Худший вариант: нулевая полнота (ни один релевантный документ не найден: $k_r = 0$) и максимальный шум (все, что выделено, не соответствует запросу: $k_g = 1$).
5. $D_i^0 \cap D_i \neq \emptyset \ \& \ D_i^0 \not\subset D_i \ \& \ D_i \not\subset D_i^0 \ \& \ D_i^0 \neq D_i$. Имеют место и неполнота ($0 \leq k_r < 1$), и шум ($0 \leq k_g < 1$).

Определим коэффициенты полноты и шума:

$$k_f = \lim_{k \rightarrow \infty} \frac{1}{k} \sum_{i=1}^k \frac{|D_i \cap D_i^0|}{|D_i|}, \quad k_g = \lim_{k \rightarrow \infty} \frac{1}{k} \sum_{i=1}^k \frac{|D_i \setminus D_i^0|}{|D_i|},$$

где m - достаточно большое число, чтобы по теореме о больших числах обеспечить требуемую достоверность результата эксперимента по определению k_f и k_g .

Смысл коэффициентов полноты и шума



Успешность поиска формально определяется степенью совпадения множеств D_i и D_i^0 .

Сравнение документальных, фактографических и гипертекстовых ИПС по ряду показателей

Характеристики ИПС	Виды ИПС		
	Документальные	Фактографические	Гипертекстовые
Полнота и шум	$k_{i \max} = 0,5$ $k_{\emptyset \max} = 0$	$k_{i \max} = 1$ $k_{\emptyset \max} = 0$	$k_{i \max} = 0,9 \dots 1,0$ $k_{\emptyset \max} = 0,1 \dots 0,2$
Систематизирующая информация	Поисковые образы документов, метаданные	Значения атрибутов объектов ПрО	Гипертекстовое представление документов, метаданные
Тип поискового аппарата	Информационно-поисковые языки с развитой грамматикой	Языки реляционного типа	Гипертекстовый тезаурус
Трудоемкость подготовки информационного массива	Требуется специальная лингвистическая подготовка сотрудника	Требуется высокая квалификация сотрудника	Относительно несложная подготовка по типам семантических связей
Структуры данных	Прямые и инверсные списки	Иерархические или реляционные структуры	Семантическая сеть: вершины – понятия, ребра – отношения
Математический характер критериев поиска	Логические и алгебраические выражения	Логические и алгебраические выражения	Семантические признаки
Тип собственного языка системы	Специальные информационные языки (например, Сетка-5)	Специальные языки (SQL, QBE)	ОЕЯ ПрО

Системы контекстной помощи

Системы контекстной помощи можно рассматривать как частный случай интеллектуальных гипертекстовых и естественно-языковых систем. В отличие от обычных систем помощи, навязывающих пользователю схему поиска требуемой информации, в системах контекстной помощи пользователь описывает проблему (ситуацию), а система с помощью дополнительного диалога ее конкретизирует и сама выполняет поиск относящихся к ситуации рекомендаций. Такие системы относятся к классу систем распространения знаний (Knowledge Publishing) и создаются как приложение к системам документации (например, технической документации по эксплуатации товаров).