# 1 Этапы развития сети интернет

1969 – создание сети ARPA net

1971 – разработана 1-я программа для отправки почты

1973 – сеть стала международной

1982-83 стандартизация протоколов передачи данных (с участием Постела)

1984 – была разработана DNS, появилась сеть NSF net

1983 – закрепляется название Интернет, NSFnet вытеснила ARPA net

1988 – появляется протокол IRC

Конец 90-х – появились первые реализации

1989 – Тим Бернес Ли публикует работы по HTML, HTTP, URI

1993 – появление 1-го браузера NSCA Mosaic

1995 – Всемирная сеть становится основным источником информации

1998 – папа Римский Иоанн Павел II учредил всемирный День Интернета (30 сентября)

2 Понятие интеллектуальной системы.

Под интеллектуальными системами понимают любые биологические,

искусственные или формальные системы, проявляющие способность к целенаправленному поведению. Последнее включает свойства (проявления) общения, накопления знаний, принятия решений, обучения, адаптации и т. д.

В настоящее время существует устойчивая тенденция интеллектуализации компьютеров и их программного обеспечения (ПО). Основные функции будущих компьютеров — решение задач все в большей степени невычислительного характера, в том числе логический вывод, управление базами знаний (БЗ), обеспечение интеллектуальных интерфейсов и др. Интеллектуализация компьютеров осуществляется за счет разработки как специальной аппаратуры (например, нейрокомпьютеры), так и ПО (экспертные системы, базы знаний, решатели задач и т. д.).

3 Функции интеллектуальной системы.

1. Функция представления и обработки знаний. Заключается в том, что ИС должна накапливать знания об окружающем мире, классифицировать и оценивать их с точки зрения прагматики и непротиворечивости, инициировать процессы получения новых знаний, соотносить новые знания со знаниями, хранящимися в базе знаний
2. Функция рассуждения. ИС должна быть способна формировать новые знания с помощью логического вывода и механизмов выявления закономерностей, получать обобщенное знание на основе частных знаний, логически планировать свою деятельность
3. Функция общения. ИС должна быть способна общаться с человеком на языке, близком к естественному, получать информацию через каналы, аналогичные тем, которые использует человек для восприятия окружающего мира

4 Блоки интеллектуальной системы

 интеллектуальный интерфейс обеспечивает общение с внешней средой и преобразование информации из внешнего во внутреннее представление и обратно; • подсистема логического вывода на основе анализа семантики вход­

ных сообщений и имеющихся знаний формулирует постановку задачи, осу­

ществляет поиск вариантов ее решения и выбирает из них наилучшие;

• генератор программ формирует программу решения, используя зна­

ния о методах решения задач;

• интерпретатор задач обеспечивает выполнение сгенерированных

программ;

• база знаний обеспечивает хранение и доступ к различным видам знаний, используемым интеллектуальной автоматизированной системой (ИАС) при ее функционировании.



5 Понятие гипертекста.

**Возникновение понятия гипертекста**

Впервые понятие появилось в 1945 году. Ванневар Буш высказал идею машины для просмотра и пополнения записей документов, записанных на пленке. Термин гипертекст был предложен Тедом Нельсоном в 1962 году. Впервые опубликовал в 1965 г в описаниях гипертекстовой системы для хранения теста.

Первые ГТ системы. 1-я система была предложена Дугом Энгельбартом. В процессе разработки NLS он разработал мышь, многооконный интерфейс, клавиатуру. В 1989 Тим Бернес Ли работая над внутренней сетью организации CERN предложил глобальный ГТ проект во Всемирной паутине.

Под ГТ понимается форма организации семантической информации, предусматривающая ее разделение на фрагменты, для каждого из которых заданы переходы к родственным фрагментам. Исторически первым ГТ можно назвать Библию.

6 Основные понятия гипертекстовой информационной технологии

**Основные идеи Гипертекста.**

1. Текст разбивается на фрагменты , представляющие его семантические единицы( сеты).

Между ними устанавливается связи, которые могут наделятся именами.

1. ГТ можно читать по различным траекториям, образованными связанными сетами
2. Активируемые переходы выбираются читателем, а типы имен облегчают решение задачи выбора перехода

### ***ГТ и графовые модели***

В гипертекстовом документе могут быть представлено несколько уровней детализации материала. Такие документы моделируются деревьями или сетями

В графовой модели ГТ вершины вычленены и соответствуют фрагментам текста, а ребра – возможным переходам между ними

Каждый путь на графе представляет отдельную линию прочтения текста.

Дерево – это остов графа

Сеть – это граф, в котором возможны циклы

### *ГТ информационные технологии*

Технология обработки семантической информации на использовании ГТ

Модели ГТ

- Формализованная модель ГТ

- Условно- типовая модель ГТ

Понятия ИСС

7 Аспекты применения гипертекстов в среде интернет.

Коротко поясним основные аспекты применения ГИТ в Internet. Информационные ресурсы Internet разнородны и динамичны. Их невозможно представить в виде единой БД. Гипертекст в Internet применяется с 1993 г. в рамках технологии World Wide Web (WWW) — «всемирной паутины», позволяющей перемещаться по сети гипертекстовых документов. В соответствии с протоколом передачи гипертекста HyperText Transport Protocol (HTTP) минимальной неделимой единицей данных, предназначенной для межмашинного обмена, является текст, записанный на языке разметки гипертекста HyperText Markup Language (HTML). Файл с этим текстом представляет собой гипертекстовый документ, называемый НТМL-страницей или web-страницей. HTML-страница содержит описание структуры документа, в тело которого в виде унифицированного указателя ресурса (Uniform Resource Locator — URL) могут входить ссылки на фрагменты данного документа и других документов.

Взаимосвязанная совокупность HTML-страниц, расположенных на одном web-сервере, образует web-сайт. Гипертекстовый документ, представленный на HTML, может включать не только текст, но и таблицы, фрагменты исполняемого сервером или компьютером пользователя программного кода (скрипты, апплеты), а также ссылки на цифровые объекты (графические изображения, звук, видео, анимацию и др.). Отметим, что возможности HTML как языка описания данных

выходят за рамки только лишь включения в документ гипертекстовой разметки. В частности, язык HTML позволяет:

• определять структуру документа (заголовки и области различных уровней);

• представлять собственно содержимое документа;

• устанавливать оформление содержимого (способ представления информации — отступы, шрифты, цвета, выравнивание, параметры таблиц и т. д.);

• задавать ссылки для вставки внешних компонентов — рисунков, элементов пользовательского интерфейса, программных объектов и др. (их вставка или активация происходит на этапе загрузки страницы);

• включать в документ фрагменты программного кода (скрипты);

• определять гиперссылки, ассоциируемые с различными информационными элементами документа для организации переходов и вызова функций.

Логически единая система HTML-страниц может быть физически рассредоточена по сети. Система URL позволяет как размещать, так и собирать ресурсы, на которые ссылается ГТ.

8 Формализованная модель гипертекста.

В формализованной модели гт ИСС соответствует информационному объекту, содержание которого характеризуется смысловым единством и логической целостностью.

В формализованной модели описывается кортеж (x0, x1, ……, x11)

Где

 х0 –имя ИСС

Х1– заголовок

Х2 – аннотация

Х3 – точка входа в ИСС

Х4- множество текстовых фрагментов исс

Х5 – множество цифровых фрагментов

Х6 – множество программных фрагментов

Х7 – справка по ИСС

Х8 - признак быстрого просмотра исс

Х9 – признак детального просмотра исс

Х10 - список гиперссылок внутри исс

Х11 – список гиперссылок между исс

 Два слоя гипертекстовой модели: 1-й слой представляет отображение на экране содержания документа, в котором гиперссылки по умолчанию выделены цветом или подчеркиванием и изменением шрифта.

Во втором слое хранятся адреса переходов (идентификаторы исс, метки их фрагментов)

Недостаток форм. Модели – отсутствие в ней возможности явного определения гипрессылок.

Достоинство условно типовой– все гиперссылки имеют явно указанный тип.

9 Условно-типовая модель гипертекста.

Состав условно – типовой модели

Обязательным компонентом является тезаурус – автоматизированный словарь, отображающий семантические отношения между лексическими единицами дескрипторного информационно – поискового языка и предназначен для поиска слов по их смысловому содержанию

Каждый термин сопровождается его текстовой характеристикой. Тезаурус позволяет пользователю Г Т системы уточнять как содержание (смысл) , так и объем интересующего его термина.

Для упрощения работы с гипертекстом и повышения эффективности поиска по нему в условно – типовую модель включается список главных тем и указатели.

***Список главных тем*** – сегменты, соответствующие более или менее независимым частям (срезам или аспектам) предметной области, он отражает самое общее представление о тематике.

***Указатель*** – упорядоченная установленным образом последовательность информационных объектов (понятий, выражений) ссылающихся на информационно – справочную страницу, в которой эти объекты упоминаются.

В зависимости от характера объектов, указатели подразделяются на:

Предментые, библиографические, именные, событийные, и т.д.

По принципу упорядочивание различают: алфавитные, систематические, хронологические.

В ГТ могут использоваться несколько типов указателей

В лингвистике выделяют около 200 семантических типов отношений

10 наиболее употребляемых : синоним, род- вид, вид-род, часть-целое, целое-часть, процесс-надпроцесс, процесс-подпроцесс, причина-следствие, следствие- причина, ассоциация

Графовой интерпретацией условно-типовой модели является семантическая сеть

***ИСС для условно -типовой модели ГТ***

Имя заголовок собственно текст локальный справочный аппарат исс

Локальный справочный аппарат исс

Список ссылок на исс, связанных с данной различными типами отношений. При этом ссылки относятся только к ближайшим родственникам (связанным например через 1 -0 отношение)

Способы организации справочного аппарата

1. В виде списка
2. Ссылки внедряются в текст
3. Часть ссылок помещается после заголовка статьи в виде списка, оставшаяся часть – в самом тексте.

10 Инструментальные средства для создания гипертекста

Microsoft Win Help

Html – позволяют создавать контекстно-зависимые исс

Hyper ref – предназначена для построения электронных ГТ изданий большого объема. Разработана в МЭИ (ТУ) под руководством Тихонова

АСФОГ – создана в МЭСИ и предназначена для моделирования экономических аспектов и процессов на основе представления информационного фонда

Особенностью hyper ref является то, что она позволяет вводить информацию не в ручном, а в потоковом виде. В ней предусмотрены средства, присущие фактографическим и полнотекстовым БД : это словари ключевых слов, оглавления, средства выполнения сложных запросов, средства автоматической индексации текста.

Особенности АСФОГ - система предназначена для использования при моделировании слабо структурированных предметных областей.

 ГТ ИПС

11 Понятие информационного поиска

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

Поисковый образ – признаки документа, отражающие его содержание в ИПС.
Поисковое предписание – признаки запроса к ИПС.

Индексирование – процедура перевода документа и запроса в форму представления принятую в ИПС.

Релевантость – критерий смыслового соответствия при сопоставлении поискового образа и поискового предписания.

Классы ИПС:

* Документальные
* Фактогорафические
* Гипертекстовые

12 понятие ипс

13 Классы информационно-поисковых систем

Документальные ипс хранят и выдают сведения о документах основное содержимое которых представлено в виде связного текста на ЕЯ.

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

Базовые определения документальных ипс

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. Ключевые слова ( дескрипторы) – ряд слов или словосочетаний, количество которых значительно меньше слов реферата, и которые достаточно точно характеризуют описание.

При вводе в ИПС нового объекта формируется реферат, на основании которого формируются дескрипторы, которые автоматически включаются в словарь дескрипторов.

Каждому дескриптору присваивается номер, который называется индексом дескриптора.

Совокупность индексов соответстувет полному набору дескрипторов реферата, составляет его поисковый образ. Новый поисковый образ снабжается уникальным идентификатором, и включается в массив поисковых образов. Тем же идентификатором помечается новый реферат, заносимый в массив рефератов.

***Алгоритм поиска в документальных ипс***

Анализ запроса на ЕЯ

Выделяются дескрипторы, их совокупность образует поисковое предписание.

Определение релевантности: сопоставление поискового предписания и поискового образа, в результате чего определяется их релевантность

Если поисковые образ и предписание релевантны

То из поискового образа извлекается идентификатор реферата.

Ответом на запрос ялвяется множество рефератов, соответсвтующих отобранным в процессе поиска идентификаторам

Если поисковый образ и предписание не релевантны

Поиск завершается

В целях ускорения поиска для каждого дескриптора в словаре дескрипторов указывается список идентификаторов рефератов, в которых он встречается. Такая информационная структура называется индексом. Документальный поиск относится к числу наиболее сложных информационных процессов, т.к связан с проблемой оценивания соответствия документа и запроса. Развитием поиска по дескрипторам является полнотекстовый поиск.
***Фактографические ипс***

В них хранятся не документы, а собственно сведения ( факты ) об объеме предметной области.

Подобные БД реализуются в частности на основе реляционных БД

С точки зрения релевантности результатов поиска запросу фактографический поиск, в отличие от документального, является точным и полным.

Гипертекстовые ИПС

В них кроме содержимого документов отражается их семантическая структура . поэтому по глубине формализации ГТИПС занимают промежуточные положение между документальными и фактографическими ипс.

Еще одно направление развития технологии документальных ИПС связано со структуризацией и унификацией сведений о документах. Такие сведения по отношению к исходным документам играют роль метаданных Метаданные – иформация, характеризующая какую-либо другую информацию.

Поиск по метаданным позволяет сблизить технологии документальных и фактографических поисковых систем. В настоящее время хранилище метаданных реализуют на основе реляционных и xml- ориентированных БД

Методы информационного поиска в ипс



14 Оценки качества информационного поиска

D — множество документов в информационном хранилище, $di\in D$ — i-й документ, $D\_{j}$^ 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({$ di$, dj)≥ 0. При г = 0 документы $di$ и dj эквивалентны по смыслу. Для семантически несопоставимых документов г не определена. Также введем оценки ряда важных свойств документов: S= $(S\_{1},S\_{2}, ..., S\_{k})$, к>0.

Пусть оценка каждого свойства S выражается действительным числом, принадлежащим некоторому интервалу. Для определенности примем, что чем больше значение, тем важнее для пользователя документ.

Поисковый запрос может рассматриваться как виртуальный документ z.

В идеальном случае $(r(z, d\_{i} = 0)$ ему точно соответствует документ $d\_{i}$.

Используя введенные обозначения, определим следующие виды поиска.

1. Найти ($D\_{j}$ $⊆$ D)| r(z, $d\_{i}\in $ $D\_{j}$) $\rightarrow $ min. Если $D\_{j}$= $⊘$, то в D нет документов, релевантных запросу.

При $|D\_{j}|$ =1 есть единственный подходящий документ. Если же |$D\_{j}$| > 1, то таких документов несколько.

2. Найти $\left(D\_{j} ⊆D\right)| r(z, d\_{i}\in D\_{j})$ $\leq $Δ, где Δ — оценка наибольшего допустимого расхождения смыслов запроса и искомых документов.

3. Найти ($D\_{j} ⊆D$)| $S\_{f}(d\_{i}е D\_{j}) \rightarrow $max. Результатом поиска служит подмножество документов, которым приписана наибольшая оценка важности у^го свойства. Обобщением этого варианта является векторный поиск, учитывающий оценки нескольких свойств.

4. Комбинированный поиск: найти ($D\_{j}⊆D$)|$ r(z, d\_{i}G D\_{j}) \leq $Δ $\& S\_{f}(d\_{i}e D\_{j})$ $\rightarrow $max.

Интеллектуальные возможности Ипс обусловливаются способами задания и вычисления смысловой близости r и вектора свойств документа S.

Эффективность информационного поиска документов, обеспечиваемая ИПС, оценивается по информационной полнтоте и информационному шуму. Названные показатели выражаются коэффициентом полноты kп и кш соответственно

Кп и кш прнимают значения от 0 до 1

Начальные условия

ИПС предъявляет итый запрос ИПС содержит Di релевантное запросу в результате поиска содержится Dio. Возможны следующие варианты результатов поиска

1. Идеальный вариант Dio =Di, kп=1, кш= 0
2. Di вкл в Di, 0≤кп≤1, кш=0
3. Di0включает Di , кп= 1,0≤ кш≤1
4. Di0 пересечение Di0!=пустое множество & Di != пустое множество Di!=пустое множество кп = 0, кш= 1
5. Di0 пересечение Di =пустое множество & Di0 $невключается в$ D & Di невкл в Di0 & Di0!=Di

0<Кп<1 ,кш =

* 
* 

Эффективность информационного поиска

* E1 = 2limэффективность
*  обобщенные комплексный показатель
* β – параметр, отражающий предпочтения пользователя ипс одному из показателй эффективности наз. Др
* при β=1 полнота и важность являются одинаково важными. на интервале [1,0) Приоритет имеет точность , на интервале [1,$\infty $) приоритет имеет полнота

частный случай

при β=1 мера Ван Ризбергена = интегрируемый показатель эффективность

β=0 мера = коэффициент точности

β= коэф. полноты

* 

15 Сравнение информационно поисковых систем.



16 Виды источников знаний.

Существуют два класса источников знаний:

• эксперты (специалисты в ПрО, для которой формируется ГТ);

• текстовые документы на ЕЯ.

Соответственно методы извлечения знаний подразделяются на два больших класса:

1) приобретение знаний от экспертов (коммуникативные методы);

2) обработка документов (текстологические методы).

Первый класс методов извлечения знаний имеет следующую структуру.

1.1. Пассивные методы.

1.1.1. Наблюдение за работой эксперта. Инженер по знаниям наблюдает за экспертом, который выполняет или имитирует выполнение своей профессиональной деятельности. Эксперт может комментировать совершаемые им действия. В ходе процесса ведется протокол (на бумаге, аудио-

или видеоносителе).

1.1.2. Запись и анализ лекций.

1.1.3. Запись и анализ вербальных отчетов. Как и в методе 1.1.1, эксперт выполняет или имитирует выполнение своей профессиональной деятельности. Отличие заключается в том, что на каждом ее шаге он объясняет принимаемые им решения, рассуждая вслух (почему совершается именно

это, а не иное действие; как было получено данное решение и т. п.). Вербальный отчет («мысли вслух») фиксируется на бумаге или аудионосителе и впоследствии анализируется инженером по знаниям.

17 Коммуникативные методы извлечения знаний

1.1. Пассивные методы.

1.1.1. Наблюдение за работой эксперта. Инженер по знаниям наблюдает за экспертом, который выполняет или имитирует выполнение своей профессиональной деятельности. Эксперт может комментировать совершаемые им действия. В ходе процесса ведется протокол (на бумаге, аудио-

или видеоносителе).

1.1.2. Запись и анализ лекций.

1.1.3. Запись и анализ вербальных отчетов. Как и в методе 1.1.1, эксперт выполняет или имитирует выполнение своей профессиональной деятельности. Отличие заключается в том, что на каждом ее шаге он объясняет принимаемые им решения, рассуждая вслух (почему совершается именно

это, а не иное действие; как было получено данное решение и т. п.). Вербальный отчет («мысли вслух») фиксируется на бумаге или аудионосителе и впоследствии анализируется инженером по знаниям.

1.2. Активные методы.

1.2.1. Работа с группой экспертов.

1.2.1.1. Метод «мозгового штурма». Этот метод является одним из наиболее известных и широко применяемых. Его цель — активизация творческого мышления за счет запрета критики высказываемых идей. Для проведения «мозгового штурма» формируется группа экспертов. Членам группы предлагается высказывать любые идеи, связанные с решением определенной проблемы. Выступления протоколируются. Обсуждение и критика идей исключаются. Последующий анализ и оценивание предложенных идей, как правило, выполняют эксперты, не участвовавшие в «мозговом штурме».

1.2.1.2. Метод «круглого стола». Метод заключается в организации обсуждения некоторой проблемы группой экспертов, наделенных равными правами. На первом этапе эксперты выступают по очереди, на втором проводится свободная дискуссия. Содержание обсуждения записывается на аудионоситель и впоследствии анализируется инженером по знаниям.

3. Автоматизация работы со знаниями, представленными в текстовом виде

1.2.1.3. Ролевые игры. В рамках рассматриваемой проблемной ситуации каждому эксперту приписывается определенная роль (тип действующего лица в этой ситуации). Игра заключается в имитации совместной деятельности, направленной на разрешение проблемы.

1.2.2. Индивидуальная работа с экспертом.

1.2.2.1. Анкетирование.

1.2.2.2. Интервьюирование.

1.2.2.3. Свободный диалог. Суть свободного диалога - беседа инженера по знаниям с экспертом, для которой заранее не составляется план интервью или перечень вопросов.

1.2.2.4. Исследовательская игра с одним экспертом. В игре участвуют эксперт и инженер по знаниям. Последний может играть одну из ролей в рамках рассматриваемой проблемной ситуации.

18 Текстологические методы

Структура второго класса методов извлечения знаний приведена ниже.

2.1. Обработка текстов на ОЕЯ.

2.1.1. Анализ специализированной документации.

2.1.2. Анализ специализированных инструктивных и нормативных материалов (должностных и производственных инструкций, методик и др.).

2.2. Обработка текстов на ЕЯ.

2.2.1. Анализ учебной литературы.

2.2.2. Анализ научной и научно-практической литературы.

2.2.3. Анализ периодических изданий.

2.2.4. Анализ технической документации.

19 Автоматизация построения гипертекста

 Автоматизированное извлечение знаний из текста



1. Выделяет в исходном тексте слово и фразы и

 Проверяет выполнение принятиых ограничений

1. Блок морфологического анализа

Выделяет из слова неизменные части и приписывает словам ряд грамматических характеристик.

1. Программная реализация проедпроцессора и блока морфологического анализа обычно не вызывает трудности, за исключением для предредактора – сложноподчиненные предложения, для морфологического анализа – случаи морфологической амонимии.
2. Блок синтаксического анализа – строится дерево синтаксического разбора, используя базу синтаксических правил. В рамках этого блока решается проблема морфологической амонимии. Реализация блока как правило не вызывает трудностей.
3. Цель семантического анализа состоит в определении для каждого слова и фразы некоторых смысловых характеристик. Проблема в реализации блока возникает из-за семантической неоднозначности слов. Для снятия этой неоднозначности используются тезаурусные статьи.
4. Семантический анализ . трудности реализации этого этапа связаны с большими размерами требуемых семантических сетей и многовариантностью анализа.
5. Перевод анализируемого текста во внутреннее представление. Как правило для этих целей используется семантическая сеть. Не вызывает трудностей.
6. Внутреннее представление является основой для реализации понимания ЕЯ- текста.

20 Место гипертекстовой технологии среди технологий искусственного интеллекта

Гипертекст расширяет возможности человека, связанные с поиском и обработкой информации, за счет установления ассоциаций, построения обобщений, формирования целостного представления о содержании документа и т. д.

В настоящее время существует тенденция интеграции гипертекстовых ИС со специализированными пакетами прикладных программ. При этом возникают гибридные ИС, предназначенные для решения различных классов трудноформализуемых задач. В ряде источников гипертекстовые ИС рассматриваются как представители систем, доставляющих

Знания

21Проблема понимания текста на естественном языке.

В настоящее время не разработано теоретических подходов, которые позволяют решать проблему понимания в полном объеме.

Понимание – многоуровневый процесс. На 1-м уровне все сведения о содержании рассматриваемого текста извлекаются в результате его анализа без привлечения дополнительных знаний. На 2-м уровне с помощью процедур логического дополнения осуществляется доопределение временной, пространственной и причинно-следственной структуры событий. На 3-м уровне сформированному содержанию текста добавляется информация релевантная этому содержанию и известная системе. На 4-м уровне к нему присоединяются сведения, полученные из БЗ. Связанные с текстом только отношением ассоциации. На 5-м из текста извлекается его прагматическое содержание.

22 Компьютерные методы поиска в тексте

Методы поиска в тексте, используемые человеком, представлены следующими формами:

• поиск «сверху» (по оглавлению с аннотациями глав и, возможно, менее крупных разделов);

• поиск «снизу» (с помощью различных указателей);

• поиск с помощью гипертекстовых связей (перекрестных ссылок);

• полнотекстовый поиск путем просмотра всего текста.

Компьютерные методы поиска реализуются в ИПС, БД, БЗ и поисковых машинах Internet.

В информационно-поисковых системах применяются следующие методы поиска:

1) индексирование текстов и поиск по ключевым словам (по индексу);

2) поиск, включающий морфологический разбор и отождествление различных грамматических форм слов;

3) поиск с ранжированием документов по степени релевантности запросу;

4) использование формальных поисковых языков;

5) комплексные методы.

В технологиях БД и БЗ наряду с перечисленными применяются следующие методы поиска:

• использование формальных языков запросов, позволяющих описывать условия совместного вхождения ключевых слов в документ (это направление представляют SQL-подобные языки);

• методы семантического анализа текста.

Средства автоматического извлечения знаний из текстовых ресурсов Internet реализуются в поисковых машинах. При этом различают:

1) методы итеративного поиска;

2) методы поиска по выборке;

3) методы, использующие каталоги (рубрикаторы и классификаторы, организующие множество документов в деревья или лес); семантические методы поиска, использующие подходы ИИ. 23 Каталог ресурсов Интернет

Для поиска информации в Internet служат различные классы поисковых средств:

• каталоги (directories);

• подборки ссылок (bookmarks);

• поисковые машины (search engines);

• БД адресов электронной почты (email addresses databases);

• средства поиска в архивах Gopher (Gopher archives);

• системы поиска файлов (FTP search);

• системы поиска новостей (usenet news).

Каталог ресурсов Internet — постоянно обновляемая и пополняемая система ссылок на ресурсы, распределенные по иерархической структуре категорий. На верхнем уровне каталога представлены самые общие категории (рубрики), например, «наука», «бизнес», «развлечения» и т. д. На нижележащих уровнях эти рубрики декомпозируются на подчиненные рубрики,

имеющие более частный характер. На нижнем уровне каталога указываются ссылки на конкретные ресурсы Internet (сайты и web-страницы), снабженные краткими описаниями

их содержимого.

Ниже перечислены некоторые универсальные каталоги ресурсов Intemet:

• Yahoo! msn, AOL (<http://search.aol.com>);

24 Типовая поисковая машина.

Поисковые машины (или поисковые системы) позволяют находить

ресурсы Internet непосредственно по их текстовому содержимому. Функционирование поисковой машины включает два базовых процесса: 1) индексирование ресурсов Internet (автоматическое построение и обновление индекса); 2) поиск по индексу по запросам пользователей.



Ее главными компонентами являются:

• программный агент, «перемещающийся» по сети и индексирующий

ресурсы (web-страницы);

• БД (индекс), содержащая информацию, собираемую агентом;

• программа поиска, применяемая пользователями для поиска ин­

формации в БД.

25 Алгоритм работы поисковой машины

Алгоритм работы поисковой машины при индексации

1. Адреса web-узлов, включаемые в обрабатываемую область, определяются по гиперссылкам, ведущим из страниц данного web-узла. При этом используются различные модификации волнового алгоритма (например, с вычислением профилей узлов).

2. Агент либо переходит к индексированию очередного web-узла из сформированного списка, либо выполняет так называемое зеркалирование (дублирование) его содержимого на свой web-узел.

3. Производится собственно индексирование. Оно может быть полнотекстовым (обрабатывается весь текст) и неполнотекстовым (обрабатываются наиболее значимые части текста: заголовки, названия, ключевые поля, начальные слова разделов и т. д.).

4. Полученные данные о ключевых словах добавляются в БД.

5. Если был сделан зеркальный дубль, он стирается.

6. Пункты 2—5 повторяются для каждого адреса, полученного в п. 1. Изложенный алгоритм соответствует некоторой канонической структуре поисковой машины. Конкретные их реализации различаются по многим параметрам:

• поддержке простого и сложного поиска;

• учету различий строчных и прописных символов;

• возможности поиска по частям слов и словосочетаниям;

• поддержке обработки запросов, содержащих логические операторы

И, ИЛИ, НЕ;

• использованию специальных языков поиска информации, значительно сокращающих его время (к сожалению, такие языки не стандартизованы, поэтому в разных поисковых машинах реализуются разные поисковые языки). Применение поисковых машин для поиска в Internet эффективно, если пользователь представляет, какие ключевые слова характеризуют нужные

ресурсы.

Дополнительные возможности предоставляет режим расширенного поиска, в котором можно задавать правила поиска. Часто это значительно увеличивает вероятность нахождения требуемой информации.

Агент - самый интеллектуальный из компонентов поисковой машины.

Он обладает автономностью, имеет блоки навигации, управляющие «перемещением» по сети, и механизмы индексации, основанные на некоторой базе правил. Агенты реализуются как простые программные системы, запрашивающие информацию с узлов Internet. Физически по сети агенты не перемещаются. Они индексируют полученные страницы и заносят результа­

ты в БД.

Поисковые механизмы отличаются разнообразием. Некоторые агенты следуют по каждой ссылке на каждой найденной странице и затем, в свою очередь, исследуют каждую ссылку на новой странице и т. д. Как правило, агенты игнорируют ссылки к графическим и мультимедийным файлам, файлам с данными (например, архивам), БД и др. Ряд агентов просматривают страницы с учетом их популярности.

Одной из проблем является реализация алгоритма перемещения (навигации) по сети. Учитывая, что большинство web-серверов организовано иерархически, перемещение вширь по ссылкам от исходной вершины при ограниченной глубине вложенности с большей вероятностью приводит к

нахождению документов с высоким уровнем релевантности, чем при перемещении в глубину. Поскольку это подтверждается статистикой работы поисковых машин, данный метод (сначала вширь, затем вглубь) принят как предпочтительный для индексирования web-ресурсов.

Разновидностями агентов являются кроулеры, «роботы» и «пауки».

Кроулеры (crawlers) просматривают заголовки страниц и возвращают поисковой машине только первую найденную ссылку. «Роботы>> проходят по ссылкам различной глубины и вложенности. «Пауки» (spiders) сообщают о содержании найденного документа, индексируют его и пересылают извлеченную информацию в БД поисковой машины.

Системой правил для всего этого сообщества автономных программ управляют администраторы поисковых машин. Они же устанавливают параметры алгоритмов определения степени релевантности документа и запроса.

 Обычно в этих алгоритмах учитываются:

• количество слов запроса в текстовом содержимом документа (т. е. в

HTML-коде);

• теги, в которых эти слова встречаются;

• местоположение искомых слов в документе;

• удельный вес слов, относительно которых определяется релевант­

ность, в общем количестве слов документа;

• время существования web-сайта;

• индекс цитируемости web-сайта и др.

26 Определение метаданных

Системы и модели метаданных

Метаданные (metadata) — это информация о документе, понимаемая ЭВМ, т. е. обладающая свойством внутренней интерпретируемости. В общем случае метаданные представляют собой информацию, характеризующую какую-либо другую информацию. Экземпляр метаданных для ИР выступает в качестве описания этого ИР.

Консорциум Meta Data Coalition определяет метаданные как описательную информацию о структуре и смысле данных, а также приложений и процессов, которые манипулируют данными

27 Понятие системы метаданных

система метаданных – представляет собой стандарт, предназначенный для формального описания категории ресурсов такой набор включает в себя набор полей(атрибоутов), свойств, элементов метаданных, характеризующих рассматриваемый объект для метаданных

К числу основных требований к системе метаданных относятся [97]:

• универсальность в рамках установленного понимания ИР как объекта систематизации;

• структурированность и формализованность метаданных, необходимые для их автоматической обработки;

• достаточная выразительность для обеспечения эффективного решения задач, требующих наличия метаданных;

• совместимость с международными стандартами и протоколами в области метаданных и информационного поиска (создание условий для интероперабельности);

• возможность задания ограничений целостности, отражающих взаимосвязи полей описания ИР;

• обеспечение возможности хранения метаданных как совместно с ИР, так и отдельно от него;

• возможность представления в метаданных сведений о создателях, правообладателях и распространителях ИР, а также отношений между ИР. На основе системы метаданных реализуются базовые технологические процессы в электронных библиотеках: навигация по каталогу ИР;

поиск ИР; ввод и организация хранения ИР, а также исключение ИР из хранилища;

управление правами доступа к ИР, включая защиту авторских прав, организацию платы за пользование ИР и др.

В настоящее время в электронных библиотеках принято выделять две основные информационные составляющие:

1) собственно база (массив) ИР;

2) хранящаяся отдельно либо выделенная функционально база метаданных для этих ИР.

Между названными составляющими существует взаимно однозначное соответствие, на основе которого организуются процессы информационного поиска, что требует обязательного формализованного представления метаданных, т. е. разработки модели метаданных.

28 Уровни представления метаданных

определяются два уровня представления:

• инфологический, фиксируемый схемой метаданных, которая отражает состав и структуру элементов данных (полей) в экземпляре метаданных, их семантику, типы значений (включая словари и классификаторы) и ограничения целостности;

• даталогический, фиксируемый форматом метаданных, который отражает способ представления (кодирования) информации.

29 Модели метаданных. RDF (Resource Description Framework).

Одной из наиболее перспективных моделей метаданных на сегодняшний день является модель RDF (Resource Description Framework), разработанная консорциумом W3C. Она определяет основные принципы представления и обработки метаданных и обеспечивает функциональную совместимость web-приложений, обменивающихся такой информацией. В RDF использованы принципы объектно-ориентированного моделирования, элементы языков HTML, SGML и XML. Синтаксис метаданных в RDF описывается на основе языка XML. Несмотря на то, что RDF была разработана в расчете на XML-платформу, она не зависит от XML. Данная модель позволяет представлять семантическую структуру XML-документов и выражать смысл этих и иных ресурсов WWW.

Описание семантики одного или нескольких ИР средствами RDF называется RDF-спецификацией. Базовыми категориями такого описания являются ИР (субъект), свойство (предикат) и значение (объект).



Для определения информационных моделей, в соответствии с которыми должны строиться конкретные RDF-спецификации, предназначены метамодель и язык RDF Schema. В их основе лежат принципы объектно-ориентированного моделирования.

По назначению выделяют четыре основных вида метаданных:

• описательные (библиографические описания PIP и описания их семантики в виде рефератов и аннотаций);

• структурные (формат, объем и структура ИР);

• административные (правообладатели, права на доступ и коррекцию ИР, сведения о пользователях, платежах и т. п.);

• идентифицирующие, служащие для однозначного представления описываемых объектов для внешнего мира и приложений.

30 Прикладные системы метаданных

Перечислим некоторые системы метаданных:

• «Дублинское ядро» (инвариантный к ПрО набор наиболее общих полей описания ИР, введенный для обеспечения глобальной интероперабельности приложений, работающих с метаданными) [106];

• MARC — предназначена для описания библиотечных ресурсов (как на бумажных, так и на электронных носителях) [105];

• GILS — предназначена для описания любых видов ИР, расширяющая MARC и базирующаяся на протоколе Z39.50;

• ONIX — предназначена для описания товаров в системах электронной коммерции;

• LOM — предназначена для описания образовательных ИР;

• IAFA/WHOIS++ — предназначена для описания сетевых ИР;

• UDDI — предназначена для описания web-сервисов;

• ESfDECS — ориентирована на системы электронной коммерции и содержащая элементы для управления правами на цифровые объекты;

• EAD — предназначена для описания архивных материалов;

• GEM — расширение «Дублинского ядра» для описания образовательных ИР;

• МЕКОФ — международный коммуникативный формат, выступающий в качестве альтернативы MARC [99-101];

• формат описания БД и машиночитаемых информационных массивов [102].

С точки зрения ориентации на виды ИР и сферы использования различают универсальные и специализированные системы метаданных. К универсальным системам относятся «Дублинское ядро» и GILS.

31 Дублинское ядро

Наиболее распространенной системой метаданных является «Дублинское ядро» (Dublin Core Metadata Element Set). Основные цели, которые ставились при ее создании, заключались в обеспечении:

• простоты формирования и поддержки метаданных;

• легко понимаемой (как человеком, так и компьютером) семантики;

• возможности представления метаданных на разных ЕЯ;

• расширяемость системы метаданных.

«Дублинское ядро» включает два уровня:

1) простое «Дублинское ядро» (Simple Dublin Core);

2) «Дублинское ядро» с квалификаторами (Qualified Dublin Core)

Первый уровень содержит 15 элементов данных, образующих три группы (табл. 4.1):

• Content (содержание ИР);

• Intellectual Property (интеллектуальная собственность);

• Instantiation (характеристики данного экземпляра ИР).

Состав элементов простого «Дублинского ядра» определен в стандарте ISO 15836:2003\*.

На втором уровне к 15 элементам добавлены два дополнительных элемента: Audience (целевая аудитория, категория пользователей) и Rights Holder (правообладатель). Кроме того, для повышения детальности и выразительности описаний на этом уровне вводятся и используются квалификаторы, уточняющие семантику элементов данных и специфицирующие источники и способы представления их значений.

Все элементы «Дублинского ядра» являются необязательными и могут повторяться. Порядок их следования в описании ИР значения не имеет. 

Пример обобщенного абстрактного описания элемента данных:

• имя — «Предметная область»;

• идентификатор — «Subject»;

• язык — «русский»;

• определение — «ПрО, к которой относится содержание ресурса»;

• обязательность — «обязательный»;

• тип данных — «строка, максимальная длина — 200 символов»;

• максимальная распространенность — «до 10 экземпляров»;

• комментарий — «Значение выражается ключевым словом, ключевой фразой или классификационным кодом, характеризующим тему ИР и выбираемым из УДК, ГРНТИ и др.».

Сформулируем ряд рекомендаций по формированию описаний ИР на основе системы метаданных «Дублинское ядро».

1. Желательно значения элементов даь^ных выбирать из распространенных словарей (словников, тезаурусов, классификаторов). Например, источником значений элемента Coverage может служить Тезаурус географических наименований (Thesaurus of Geographic Names\*).

2. При описании темы (Subject) необходимо применять тезаурусы для представления характеристик ПрО, с которыми ассоциируется ИР (предметные рубрикаторы и классификаторы). Обычно используются два типа тезаурусов: предметные и функциональные.

Предметный тезаурус систематизирует понятия ПрО и позволяет охарактеризовать содержание ИР, т. е. ответить на вопрос «О чем этот ИР?».

Функциональный тезаурус позволяет описать роль ИР в человеческой деятельности, т. е. ответить на вопрос «Для чего нужен этот ИР?».

3. Для выражения типа ИР (Туре) может использоваться следующий минимальный набор обобщенных значений:

• text (текст);

• image (изображение);

• sound (звук);

• dataset (набор данных);

• software (программа);

• interactive (интерактивная система);

• event (событие);

• physical object (физический объект).

Каждое из перечисленных значений может служить корнем иерархического классификатора, детализирующего соответствующий обобщенный тип. Например, для типа «event» подчиненными значениями могут быть:

• конференция;

• семинар;

• круглый стол;

• выставка;

• проект.

4. Для элемента «Формат» (Format) рекомендуется выбирать значения из множества типов контента MIME (Multipuфose Internet Mail Extensions) [112, 113]. Например:

• text/xml — документ на XML;

• text/plain — текст без форматирования и разметки;

• image/gif— изображение в формате GIF.

32 Концепции «семантической паутины».

Недостатки и ограничения технологий Internet первого поколения (web-1) привели к разработке консорциумом W3C концепции «семантической паутины» (Semantic Web или web-2). Она направлена на интеллектуализацию WWW и базируется на следующих основных компонентах:

• активном использовании метаданных;

• метаязыке XML;

• онтологическом подходе, позволяющем описывать термины и отношения между ними (см. § 5.4);

• модели RDF, устанавливающей способ представления значений, определенных в онтологии.

В Semantic Web также применяются:

• универсальные идентификаторы ресурсов;

• системы обработки правил логического вывода;

• стандартные протоколы Internet.

Цель реализации Semantic Web состоит в преодолении ограничений технологий web-1 с сохранением их достоинств. К числу основных положительных черт web-1 можно отнести [119]:

• открытый характер Internet — к сети можно подключиться с помощью любого стандартного оборудования и свободно распространяемых программных средств;

• демократическая организация — использование Internet не требует существенных финансовых затрат и каких-либбо административных решений;

• эффективная как для пользователей, так и для разработчиков приложений клиент-серверная архитектура WWW;

• простота языка разметки HTML, возможность представления с помощью него не только гипертекстовых, но и гипермедийных данных, наличие множества HTML-редакторов и др.

Недостатки HTML

* Представление контента

Ориентирован не на логическую а на форматную разметку контента отражающую способ его представления

* Работа с метаданными.HTML имеет слабые средства метаданного определения структуры и семантических свойств веб-стр.
* Идентификация ИР. К html ресурсам возможен только навигационный доступ по гиперссылкам. Доступ по содержанию обеспечивают специальные средства – поисковые машины.

Около 70 % ИР Internet явно не представлены в web-1, т. е. недоступны для автоматической обработки поисковыми машинами. Подобные ресурсы образуют так называемый скрытый или глубинный web {deep web) — это БД, интегрированные в web-сайты, архивы, мультимедийные файлы, а также многочисленные документы в форматах PDF, DOC, RTF, PostScript и др.

Отсутствие эффективных методов доступа к таким ИР и описывающим их метаданным затрудняет использование web-1

33 Платформа XML

Основой web-2 служит расширяемый язык разметки XML. Возникла новая технологическая платформа web-2 — платформа XML, Под ней понимается совокупность взаимосвязанных и согласованных стандартов и спецификаций, имеющих общее функциональное назначение.

За последние годы создано ядро платформы XML. В его основе лежат стандарты XML, понятие XML-документа, способы представления метаданных с помощью XML, более общий по сравнению с URL механизм идентификации ресурсов URI, протоколы передачи XML-данных XMLP и SOAP. Модель документа может отсутствовать. XML позволяет представлять как слабоструктурированные данные (документы без модели), так и структурированные данные (документы, ссылающиеся на модели).

34 Модель документа

Модель документа – логическая разметка документа, задаваемая с помощью XML, в соответствии с некоторыми шаблонами. Модель определяется с помощью с помощью языков DTD и XML schema. В первом модель называется описанием типа документа, во втором – схемой документа.

35 Верификация XML-документов

Наличие модели позволяет автоматически верифицировать XML- документ, т. е. проверять его структуру и содержимое на соответствие ей.

Выделяются два уровня верификации:

• проверка соответствия базовому синтаксису XML;

• проверка соответствия модели.

1. Ядро спецификаций платформы XML







37 Прикладные аспекты развития семантического web-а. Онтологии.

Онтологии в semantic web

Совместно используемые формальные концепции конкретных предметных областей, они дают общее представление о понятиях, информацией из которых могут обмениваться люди и приложения

Язык описания онтология используемый в семантическом вебе: DAML, OWL

38 Прикладные аспекты развития семантического web-а. Языки запросов к RDF-хранилищам

-XQL (1998)

- XML QL(1998)

- XQuery(2005)

Языки запросов к RDF-данным

-RDF-query –предложен в 1998 в проекте Sesame

SPARQL Query lang for RDF (2006)

Логический вывод в Sem w

Rule interchange formad (rif)

-формат обмена правилами

A semantic web rule lang (swrl)

* Правила вывода новых фактов

39 Прикладные аспекты развития семантического web-а. Логический вывод. Доверие и доказательство

Доверие и доказательство

Для обеспечения целостности и непротиворечивости информации, представленной в семантическом вебе важно обеспечить связь приложений сем.в. с контекстом, а так же механизмы проверки доказательства и цифровых подписей

Данный уровень практически не сертифицирован в настоящее время

40 Прикладные аспекты развития семантического web-а. Агенты и сервисы

Агенты и сервисы

Веб сервис – программная система, предоставляющая некоторую услугу и обеспечивающая взаимодействие по сети

Функционально веб –сервис может являться агентом, а может быть обычной программой.

41 Подходы и средства практической реализации семантического web-а

Практическая реализация идей семантического веба

Проекты

Проект по автоматическому созданию rdf-описаний и хранилища метеданных, создаваемых на базе open directory поисковыми механизмами google

Проект консорциума W3C SWARD europe который занимается проблемой связи хранилищ семантических данных с использованием реляционных систем БД

Системы разработки приложений

Jena framework (java)

Drive rdf parser (c#)

42 Микроформатный подход.

Микроформаты – способ семантически различать сведения о разнообразных сущностях на веб страницах, используя стандарты HTML (xHTML)

Наиболее известные микроформаты

hCard – для публикации контактных данных людей, компаний, организаций

hCalendar – микроформат для представления семантической информации о событии в форме календаря

hAtom – ленты новостей

xFN – специальные взаимоотношения

Языки запросов к БЗ и языки обработки БЗ

Языки запросов

SQL –подобные языки( SPARQL, RDQL, TMQL, Squish)

Не sql-подобные языки (versa, errol, r-device)

Языки обработки БЗ

(внешние языки обработки (java, c#, python)

Внутренние (neno, fhat, SCP ( ostis)

Хранилища БЗ

На базе реляционных СУБД: jena, sesame, mulgara, nosql,

NEOj, sones, Info grid, hypergraphDB, allegro Graph, BigData, Dex, infinite, Graph, virtuoso, vertex DB