Министерство образования Российской Федерации
Казанский государственный технический университет им. А.Н.Туполева
Кафедра АСОИУ

Интеллектуальные информационные системы

Лекция 7

Системы обработки естественного языка. Информационно-поисковые системы

2011

ОЦЕНКА ПРОЦЕДУР ИНФОРМАЦИОННОГО ПОИСКА

$\{d_i\}$ – множество документов информационного хранилища D

$r(d_i, d_j)$ - оценка смысловой близости двух документов d_i и d_j

d_0 - некоторый воображаемый (виртуальный) документ определенного содержания.

Задача информационного поиска:

В информационном хранилище требуется отыскать:

некоторый документ d_i такой, что

$$r(d_i, d_0)=0$$

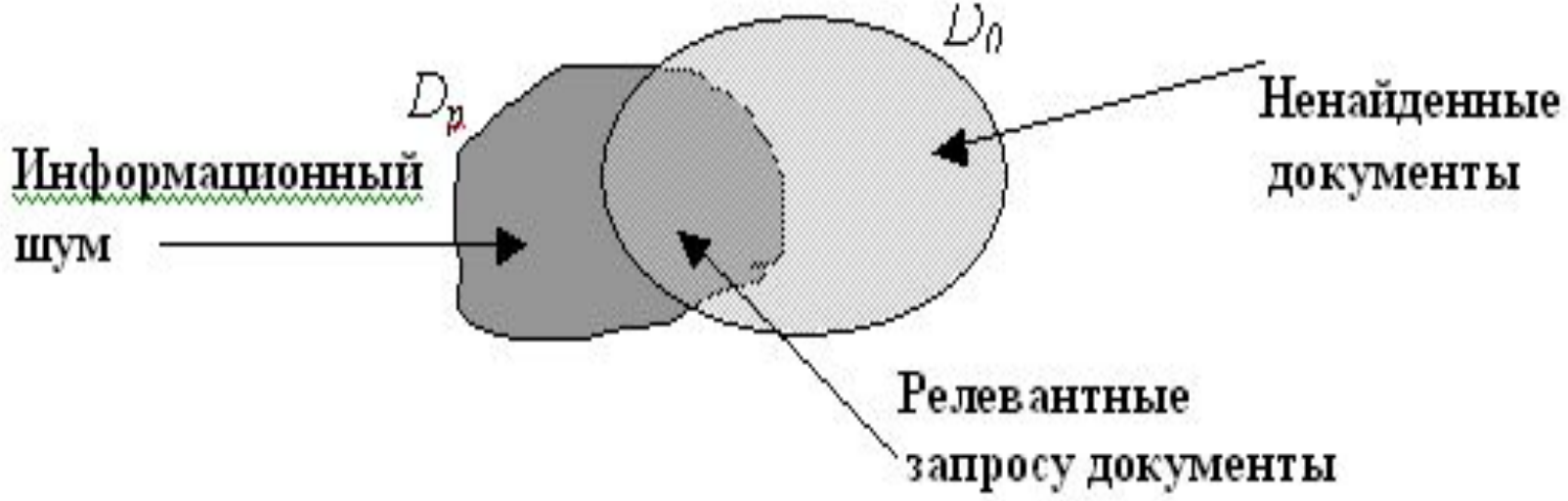
или множество документов $D_0 = \{d_i\}$, что

$r(d_i, d_0) \longrightarrow min$ для всех d_i из множества D_0.

Оценка результатов информационного поиска

Информационная полнота - все ли нужные документы найдены.

Информационный шум - отобраны и ненужные с точки зрения информационного запроса документы.



D_p - множество документов, полученных в результате выполнения поисковых процедур.

k_n - коэффициент полноты

$k_{ш'}$ - коэффициент шума

Возможные варианты результатов поиска:

1. $D_p = D_0$, т.е. найдены все адекватные смыслу запроса документы.

$$k_п = 1, \qquad k_ш = 0.$$

2. $D_p \subset D_0$ Информационный поиск является неполным:

$$0 <= k_п < 1, \qquad k_ш = 0.$$

3. $D_0 \subset D_p$, В результате поиска отобраны лишние документы (информационный шум):

$$k_п = 1, \qquad 0 <= k_ш < 1.$$

4. Пересечение D_p и D_0 не пустое, при D_p != 0 и D_0 != 0,

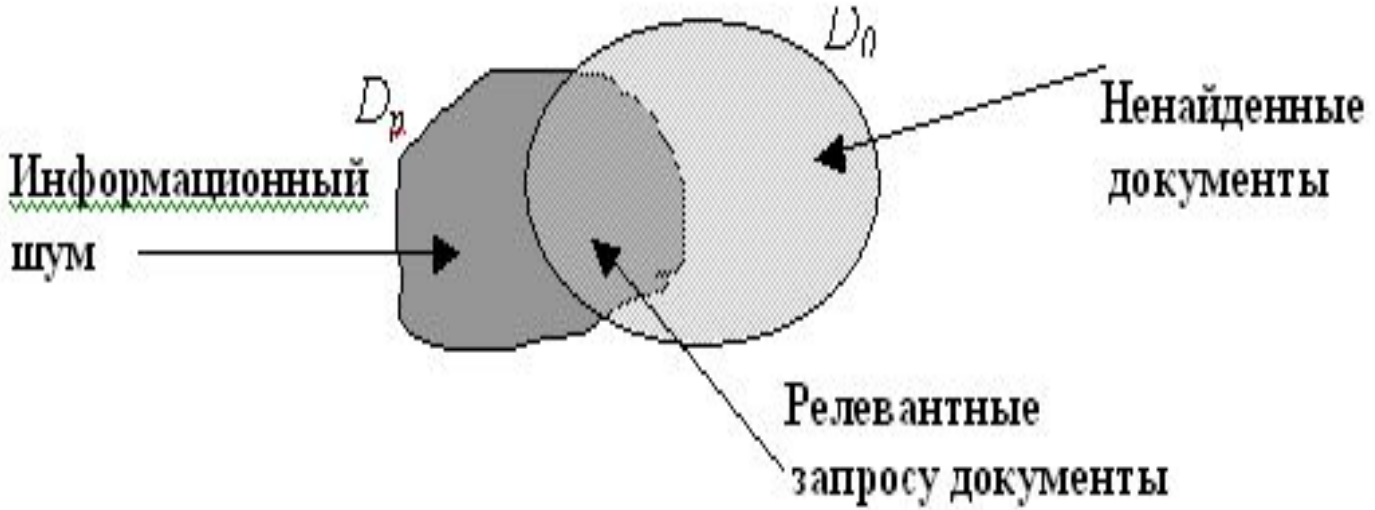
$$0 <= k_п < 1, \quad 0 <= k_ш < 1.$$

5. Пересечение D_p и D_0 пустое, при D_p != 0 и D_0 != 0,

$$k_п = 0, \qquad k_ш = 1.$$

- Вычисление k_n и $k_ш$

$$K_n = \lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k}\frac{|D_i \cap D_i^0|}{|D_i|}, \qquad K_ш = \lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k}\frac{|D_i \setminus D_i^0|}{|D_i|}$$



- Вычисление k_n и $k_ш$

$$K_n = \lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k}\frac{|D_i \cap D_i^0|}{|D_i|}, \qquad K_ш = \lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k}\frac{|D_i \setminus D_i^0|}{|D_i|}$$

- Интегральная оценка эффективности поиска

$$E_i = 2\lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k}\frac{|D_i \cap D_i^0|}{|D_i| + |D_i^0|}$$

- Вычисление $k_п$ и $k_ш$

$$K_п = \lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k}\frac{|D_i \cap D_i^0|}{|D_i|}, \qquad K_ш = \lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k}\frac{|D_i \setminus D_i^0|}{|D_i|}$$

- Интегральная оценка эффективности поиска

$$E_i = 2\lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k}\frac{|D_i \cap D_i^0|}{|D_i| + |D_i^0|}$$

- Коэффициент точности $K_\tau = 1 - k_ш$

$$E_i = \frac{2K_т K_п}{K_т + K_т}$$

Факторы, влияющие на процессы обработки информации

- Огромный объем доступной информации
- Взаимосвязи
- Высокий процент временной информации
- Неконтролируемое качество информации
- Разнотипность информации
- Избыточность
- Разнородность пользователей

Фактографические ИПС (ФИПС)

Характерные отличия::

- высокая однородность сообщений,
- фиксированный порядок следования признаков объектов

Способы поиска:

- поиск по совпадению значений для одного или нескольких признаков;
- поиск по интервалу:
- поиск, по выражению, когда используется некий логический критерий

$k_{n\,max} = 1$ при $k_{ш} = 0$

Документальные (библиографические) ИПС

Характерные отличия:

- хранение и поиск текстовых документов.

Методы поиска:

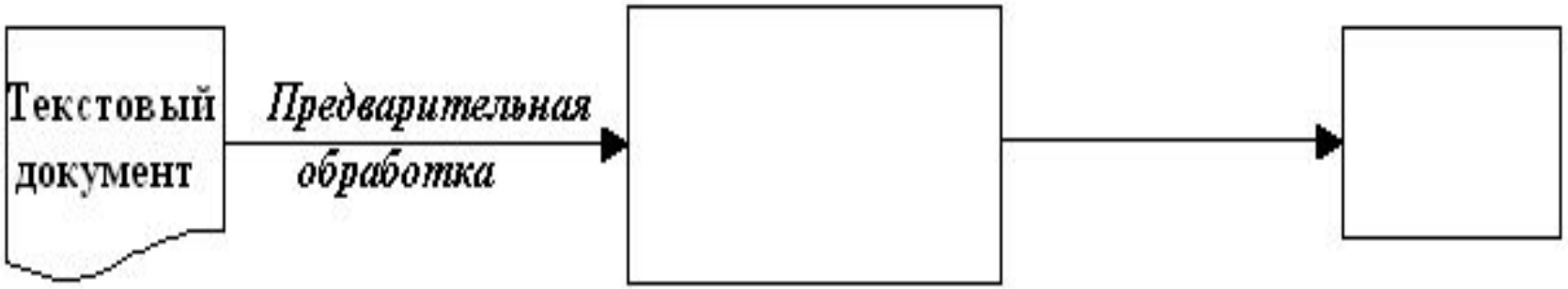
- Поиск по метаданным.
- Поиск на основе морфологического разбора.
- Поиск на основе оценок релевантности документа запросу.
- Поиск с использованием языков запросов.
- Поиск на основе семантического анализа.

$$k_{п\ max} = 0.5 \text{ при } k_{ш\ max} = 1$$

Информационный поиск в ДИПС

Дескрипторный поиск

Ввод документа в ДИПС



Информационный поиск в ДИПС

Дескрипторный поиск

Ввод документа в ДИПС

Информационный поиск в ДИПС

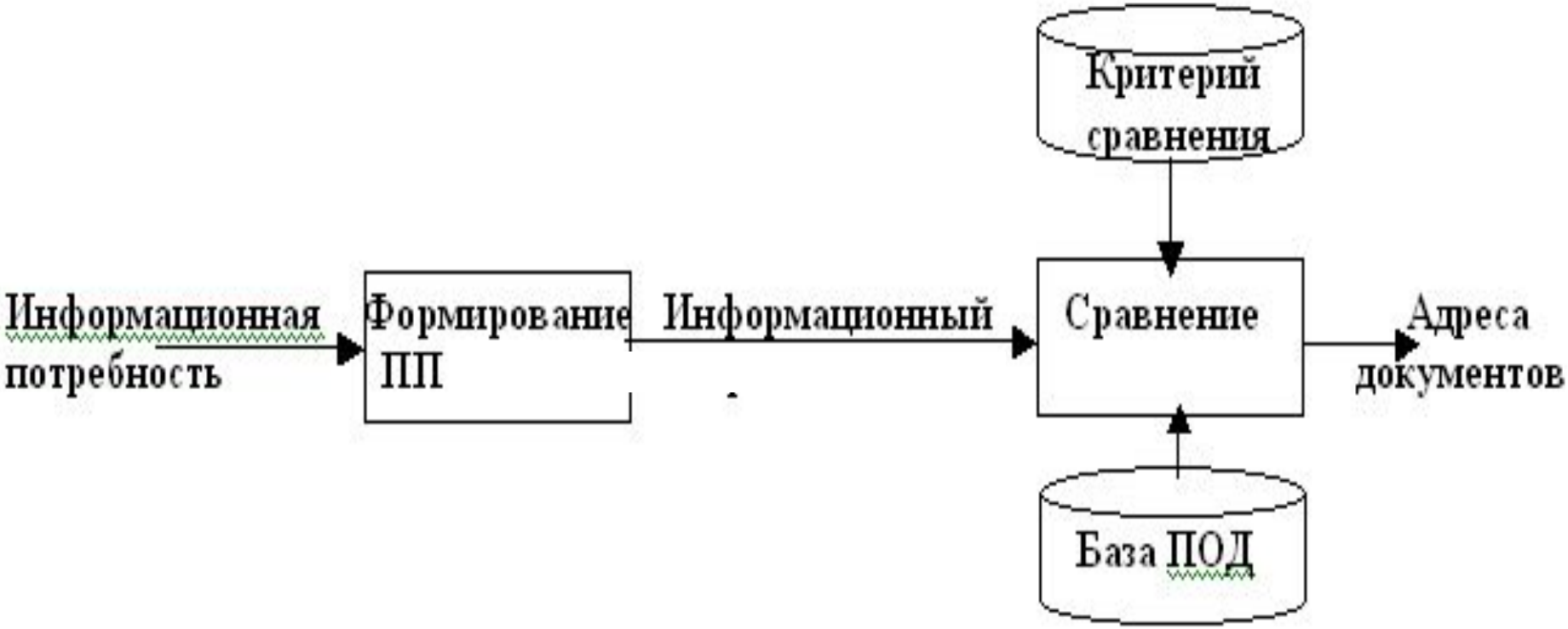
Дескрипторный поиск

Ввод документа в ДИПС

1. Формирование ПОД
2. Включение ПОД в массив ПОД.
3. Пополнение словаря дескрипторов

Информационный поиск в ДИПС

Дескрипторный поиск

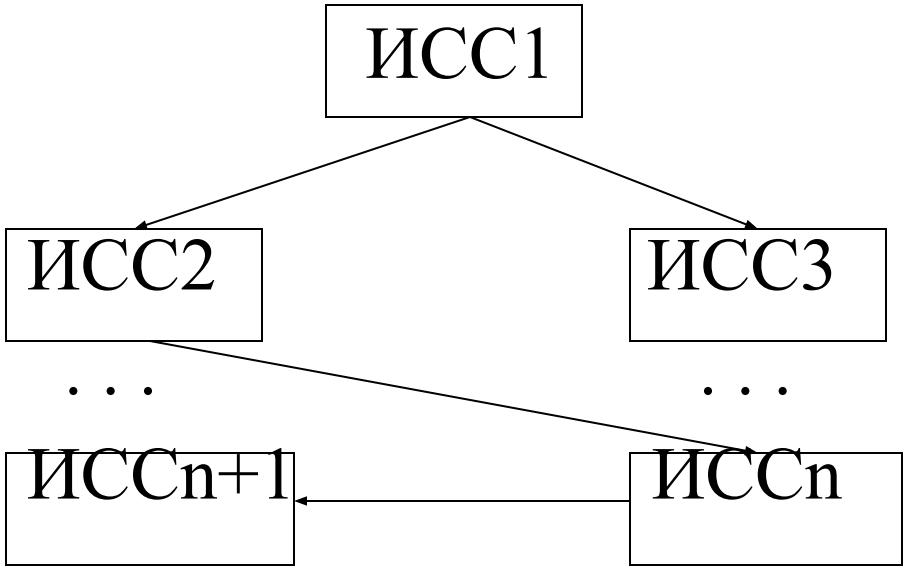


Релевантность - смысловое соответствие

ГИПЕРТЕКСТОВЫЕ ИПС

Основные идеи гипертекста:

- текст разбит на семантические единицы;
- между сетами устанавливаются связи;
- текст читается по различным траекториям.



Модель ИСС

$$(x_0,\ x_1, \ldots x_{10}, x_{11})$$

x_0 - имя ИСС,

x_1 - заголовок ИСС,

x_2 - аннотация ИСС,

x_3 - точка входа в ИСС,

x_4 - множество текстовых документов ИСС,

x_5 – множество цифровых объектов ИСС,

x_6 - множество программных объектов ИСС,

x_7 - справка по ИСС,

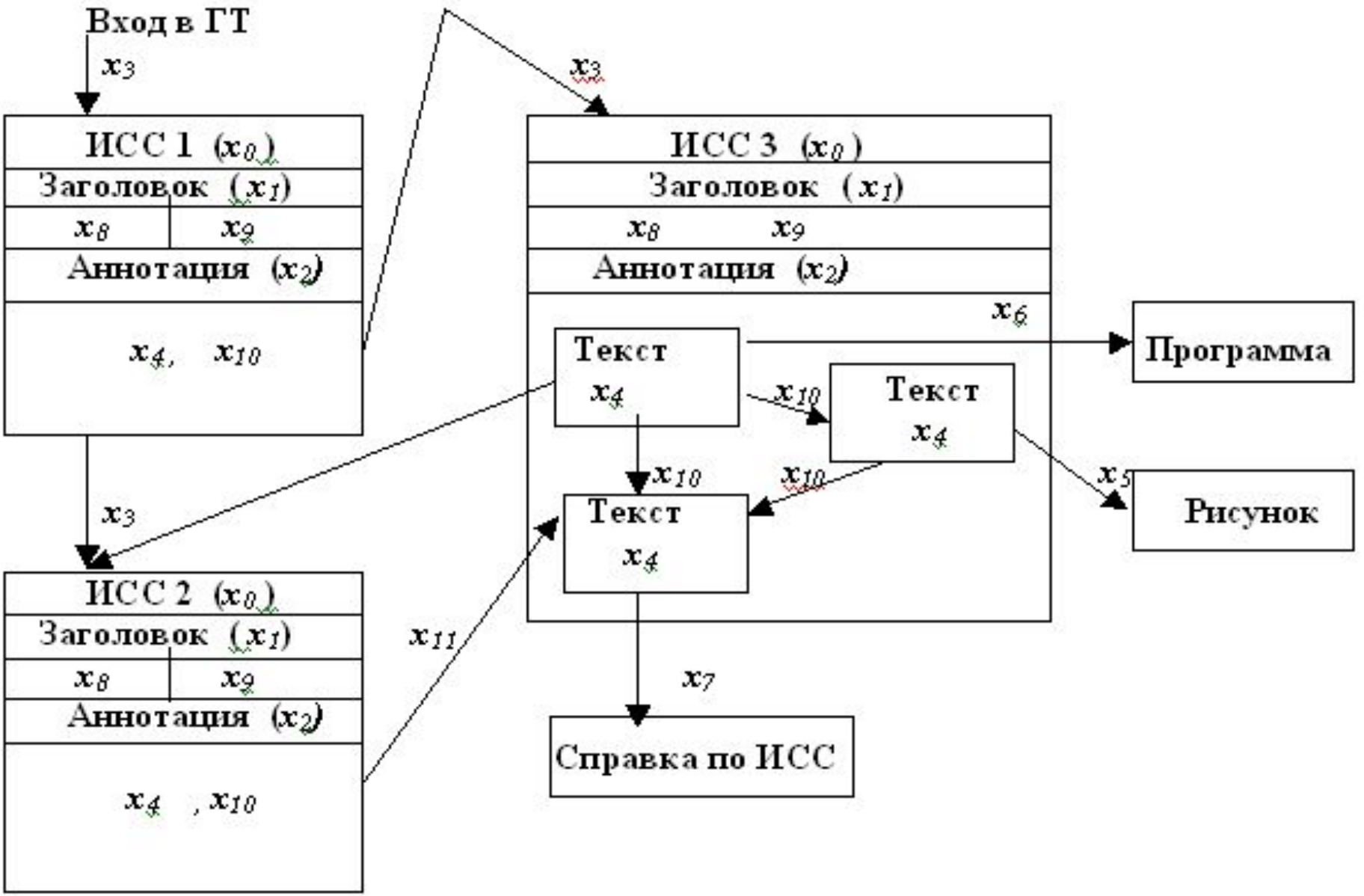
x_8 - признак ускоренного просмотра ИСС,

x_9 - признак детального просмотра,

x_{10} - список гиперссылок внутри ИСС,

x_{11} - список гиперссылок между ИСС

Структура гипертекста



ГИПС

1. **Атрибутивный поиск с помощью *SQL*-запросов, адресуемых на выполнение сопряженной СУБД.**
2. **Поиск по логической формуле, составленной из ключевых слов.**
3. **Лексический поиск на основе релевантности, оцениваемой пропорционально количеству терминов из запроса.**
4. **Ассоциативный поиск, учитывающий вхождение терминов, связанных отношением ассоциативности с терминами запроса.**
5. **Поиск ассоциаций используется для ассоциативного расширения и уточнения запроса в диалоге с пользователем.**
6. **Поиск документов по семантическому подобию.**
7. **Комбинированный поиск.**

$$k_{п\ max} = 0.9 - 1 \qquad \text{при} \quad k_{ш\ max} = 0.1 - 0.2$$

Модель поиска

- Способ представления документов
- Способ представления поисковых запросов
- Вид критерия релевантности документов

Релевантность - смысловое соответствие

Формальная релевантность.

Содержательная релевантность

Пертинентность

Простейшие модели поиска

- **Модель дескрипторного поиска**

Дескриптор - совокупность слов или словосочетаний

Дескриптор приписывается документу:

1) на основе содержания (индексирование по содержанию),

2) на основе названия (индексирование по заголовкам).

Простейшие модели поиска

- **Модель, основанная на Дублинском ядре**

  Дублинское ядро – набор метаданных, зафиксированных в спецификации определяющего стандарта.

  Образ документа D_k : $D_k = \{(N_{ik} , V_{ik})\}$

  Представление запроса: $Q = \{(N_j , V_j)\}$

  Критерий релевантности *k-го* документа: $Q \underline{\subseteq} D_k$

Модели поиска

- **Булевские модели**

  Образ документа - совокупность термов.

  $T(d_i)$ - множество термов документа d_i (**словарь документа**)

  $T = U_{i=1,\ldots,n} T(d_i)$ - словарь коллекции документов

  Представление запроса: булевское выражение.

  Критерий релевантности – истинность булевского выражения.

Модели поиска

- **Векторные модели**

Образ документа D_k : $(w_{1k}, w_{2k}, \ldots, w_{nk})$

веса термов

Например: $w_{ik} = n_{ik} / N_k$)

количество повторений i-го терма — число термов документа

Представление запроса: $(w_1, w_2, \ldots, w_n)$

Критерий релевантности вычисляется как результат операций над векторами

Модели поиска

- **Вероятностные модели**

(*PRP - Probabilistic Ranking Principle***)**

Для документа D_k определяется оценка вероятности релевантности запросу.