

# Поиск неструктурированной информации

**Информационный поиск** — процесс поиска неструктурированной документальной информации, удовлетворяющей информационные потребности.

## Виды поиска

1. **Полнотекстовый поиск** (www.yandex.ru, www.google.com)
2. **Поиск по метаданным** ( по атрибутам документа, поддерживаемым системой )
3. **Поиск изображений** (Polar Rose, Picollator)

## Методы поиска

### 1. Адресный поиск

- ✓ Наличие у документа точного адреса
- ✓ Обеспечение строгого порядка расположения документов в хранилище системы.

### 2. Семантический поиск

- ✓ Перевод содержания документов и запросов с естественного языка на информационно-поисковый язык.
- ✓ Составление поискового описания, в котором указывается дополнительное условие поиска.

### 3. Документальный поиск

- ✓ Библиотечный, направленный на нахождение первичных документов.
- ✓ Библиографический, направленный на нахождение сведений о документах

### 4. Фактографический поиск

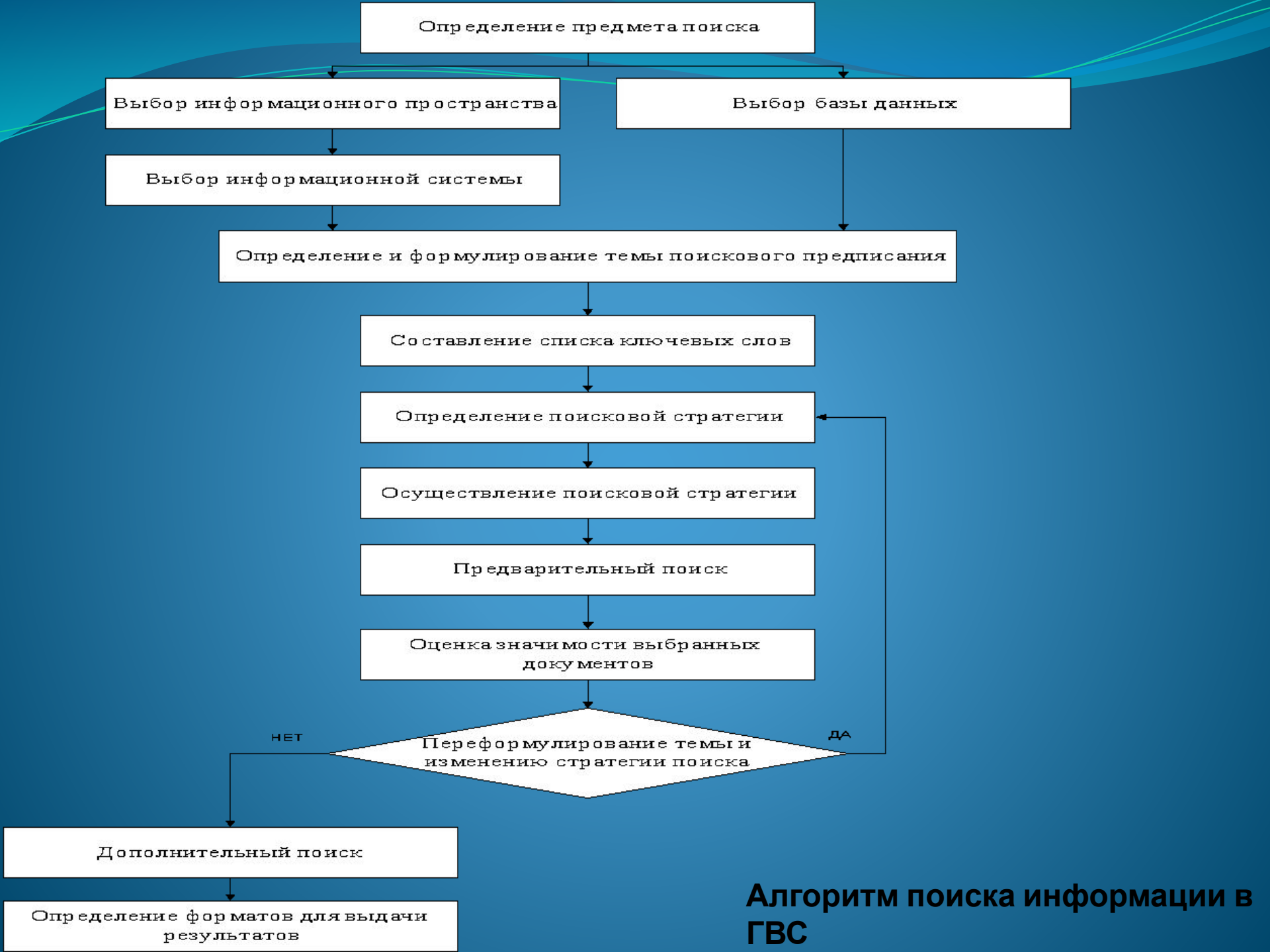
- ✓ Документально - фактографический, заключается в поиске в документах фрагментов текста, содержащих факты.
- ✓ Фактологический (описание фактов), предполагающий создание новых фактографических описаний в процессе поиска путем логической переработки найденной фактографической информации.

## Подходы к поиску информации в ГВС

В современных условиях научно-техническая и инженерная деятельность специалистов, независимо от прикладной области, немыслима без использования распределенных информационных систем глобальной вычислительной сети, предоставляющих пользователю доступ к различным знаниям.

**Успех получения информации из ГВС зависит:**

1. От знания компьютерного оборудования
2. От знания пользователя техники поиска, особенностей построения документов и баз данных в электронном виде
3. От профессионального владения предметной областью деятельности



Определение предмета поиска

Выбор информационного пространства

Выбор базы данных

Выбор информационной системы

Определение и формулирование темы поискового предписания

Составление списка ключевых слов

Определение поисковой стратегии

Осуществление поисковой стратегии

Предварительный поиск

Оценка значимости выбранных документов

НЕТ

ДА

Переформулирование темы и изменению стратегии поиска

Дополнительный поиск

Определение форматов для выдачи результатов

Алгоритм поиска информации в ГВС

# Разработка предпоисковой и поисковой стратегии

Предпоисковое взаимодействие пользователя с системой основывается на понятии информационной потребности (ИП). Границы ИП практически никогда не бывают четко определены, они стечением времени могут изменяться. Причем чем большими знаниями обладает человек, тем границы ИП шире.

**Информационная потребность** - потребность, возникающая, когда цель, стоящая перед пользователем в процессе его профессиональной деятельности либо в его социально-бытовой практике, не может быть достигнута без привлечения дополнительной информации.

Обычно используют два типа удовлетворения ИП:

**Информационный поиск, в котором четко определены границы поиска;**

- ✓ Предпоисковое взаимодействие не предусматривает ведение диалога пользователя с информационной системой
- ✓ Информационный поиск используется для получения из систем фактографической информации
- ✓ Решается одноразовым ретроспективным способом

**Информационный поиск, в котором не определены границы поиска (в дальнейшем могут изменяться).**

- ✓ Диалог пользователя с системой принципиально необходим
- ✓ Используется для получения документальной информации
- ✓ Решение осуществляется при помощи итеративного поискового процесса
- ✓ Требования к необходимости создания поискового предписания
- ✓ Составление типовых задач или определение своих задач

Основой любого ИП является **информационно-поисковый тезаурус (ИПТ)**.



**Информационно-поисковый тезаурус** - словарь дескрипторного информационно-поискового языка с зафиксированными в нем парадигматическими отношениями лексических единиц.

**Формирование ИПТ объединяет следующие этапы:**

1. Набор по специализированным текстам слов и словосочетаний, характерных для исследуемой предметной области;
2. Просмотр экспертом выявленных слов и терминологических словосочетаний;
3. Выбор из терминологических словосочетаний дескрипторов, установление синонимии и других связей между дескрипторами;
4. Пользователь должен сам сформировать свой ИПТ, на основе которого и производить формирование ПП.

В процессе формирования ПП пользователь определяет следующие декларативные компоненты:

1. перечень стандартных фраз, в которых предусмотрено включение переменных элементов в виде ключевых слов и словосочетаний, отражающих специфику предметной области;
2. словарь ключевых слов и словосочетаний (он может быть специфичным для каждой предметной области), из которой берутся переменные элементы.

# Разработка предпоисковой и поисковой стратегии

Целесообразно при формировании стандартных фраз, словарей ключевых слов и словосочетаний использовать следующие смысловые аспекты:

- ✓ описание основной темы или предмета;
- ✓ описание документа, раскрывающего или уточняющего основную тему;
- ✓ описание, посвящённое изложению (оценке) современного состояния разработок;
- ✓ цель использования;
- ✓ описание материала или объекта использования;
- ✓ описание методов (методик), приёмов и способов, использованных в процессе исследования;
- ✓ описание технических средств, инструментов или аппаратуры;
- ✓ описание конкретных результатов исследований;
- ✓ указание на область применения результатов исследований;
- ✓ описание рекомендаций использования;
- ✓ описание возможностей и перспектив развития.

*Поисковое предписание должно обеспечивать соответствие информационного запроса информационным потребностям пользователя.*

## Разработка предпоисковой и поисковой стратегии

Сложность формирования ПП определяется и тем, что к пользователю предъявляется ряд требований, специфичных для специалистов в определенной предметной области. В связи с этим, пользователь должен:

- ✓ быть квалифицированным специалистом в той области знаний, по которой производится информационный поиск;
- ✓ знать структуру и правила подготовки документов, образующих массивы информации;
- ✓ иметь знания основ математической логики и технологии поиска информации с помощью конкретного прикладного пакета программ, т.е. знать набор используемых логических операторов, оценивать их влияние на результат поиска;
- ✓ знать состав и возможности лингвистических средств, из единиц которых должно быть сформировано ПП;
- ✓ уметь правильно формулировать запрос на поиск информации.



При формулировании запроса, а как следствие и подготовки ПП необходимо:

- ✓ минимизировать текст запроса, т.е. удалить неинформативные, а при необходимости и избыточные термины;
- ✓ провести лексикографическую обработку оставшихся терминов, т.е. осуществить проверку на орфографические ошибки;
- ✓ осуществить расстановку логических элементов;
- ✓ провести пополнение логических единиц ПП нижестоящими и ассоциативными дескрипторами, т.е. провести уточнение дополнительными ключевыми словами для дальнейшего уточнения поиска.

## Поисковые системы в сети Интернет

**CRAWLER.** «путешествующий» паук, который автоматически идет по всем ссылкам, найденным на странице.

**SPIDER (паук).** браузероподобная программа, которая скачивает веб-страницы

**Indexer (индексатор):** «слепая» программа, которая анализирует веб-страницы, скаченные пауками.

**The database (база данных):** хранилище скаченных и обработанных страниц.

**Search engine results (система выдачи результатов):** извлекает результаты поиска из базы данных

## Механизм работы поисковых машин



### Основные принципы определения релевантности следующие:

1. Количество слов запроса в текстовом содержимом документа (т.е. в html-коде).
2. Тэги, в которых эти слова располагаются.
3. Местоположение искомых слов в документе.
4. Удельный вес слов, относительно которых определяется релевантность, в общем количестве слов документа.
5. Время - как долго страница находится в базе поискового сервера.
6. Индекс цитируемости - как много ссылок на данную страницу ведет с других страниц, зарегистрированных в базе поисковика.



## Метаданные, как механизм описания данных в сети интернет

- ✓ Субканальная информация об используемых данных.
- ✓ Структурированные данные, представляющие собой характеристики описываемых сущностей для целей их идентификации, поиска, оценки, управления ими.
- ✓ Набор допустимых структурированных описаний, которые доступны в явном виде и предназначение которых может помочь найти объект.
- ✓ Данные из более общей формальной системы, описывающей заданную систему данных.
- ✓ Информация о содержащейся на веб-странице информации.

## Формат метаданных

**Дублинское ядро (DCIM)** - стандарт метаданных, простой и эффективный набор для описания широчайшего диапазона сетевых ресурсов.

ГОСТ Р 7.0.10-2010 (ИСО 15836:2003)

# Инструменты, позволяющие реализовать эти стратегии

## Описание метаданных с помощью мета тегов

**Мета-теги** — HTML- или XHTML-теги, предназначенные для предоставления структурированных метаданных о веб-странице.

## Группы мета-тегов

Мета-теги разделены на две основные группы — NAME и HTTP-EQUIV.

**Группа NAME** отвечает за текстовую информацию о странице, ее авторе, а также — рекомендации для поисковых систем.

**Группа HTTP-EQUIV** фактически эквивалентны гипертекстовым заголовкам, формируют заголовок страницы и определяют его обработку. Как правило, они управляют действиями браузеров и используются для формирования информации, выдаваемой обычными заголовками.

# Инструменты, позволяющие реализовать эти стратегии

Группа NAME

**Мета-тег Author и Copyright** (идентификация автора или принадлежности документа)

**Мета-тег Description** (создании краткого описания страницы, индексация)

**Мета-тег Document-state** (Static и Dynamic, индексация)

**Мета-тег Generator** (для редактирования веб-страниц с целью саморекламы )

**Мета-тег Keywords** (Ключевые слова)

**Мета-тег Resource-type** (описывает свойство или состояние страницы)

**Мета-тег Revisit** (управлять частотой индексации документа )

**Мета-тег Robots** (разрешение индексации)

**Мета-тег Subject** (Определяет тематику документа)

**Мета-тег url** (перенаправляет робота поисковой машины по указанной ссылке)

## Теоретико-множественная модель системы поиска НТИ

Система поиска НТИ представляет собой пространство состояний  $S$  в произвольный момент времени  $t$  и включает в себя следующие основные компоненты:

$S_1$  — совокупность функций (заказов) на обработку информации, поставленных на исполнение и ожидающих (если таковая образовалась) в очереди;

$S_2$  — использование оборудования из множества рабочих мест  $A$ ;

$S_3$  — привлечение персонала из множества  $V$ .

Первая компонента описывает поток поступающих заданий на информационное обеспечение рассчитанной на предельное скопление заказов  $S_1$  на обработку информации, еще не поступивших на исполнение.

Для второй компоненты номер выполняемого задания из  $S_1$  приписывается конкретному оборудованию.

На третью компоненту возлагается задача по распределению элементарных функций обработки информации из  $S_1$  между оборудованием  $S_2$  и персоналом  $S_3$ .



## Теоретико-множественная модель системы поиска НИИ

Обработка информации, необходимой для информационной поддержки инновационной деятельности наукоемкого промышленного предприятия, включает следующие элементарные функции:

$\alpha_1(\mu, \omega)$  – прием задания заказчика (оператора)  $\mu$  на обработку информации с объекта  $\omega$ ;

$\alpha_2(b, g)$  – доступ к информационному ресурсу  $b$  с целью поиска по заданным в заказе признакам информационного объекта  $g$ ;

$\alpha_3(g, \Pi_A)$  – поиск и обработка информационного объекта с признаками  $g$  по варианту сложности (уровню аналитической нагрузки)  $\Pi_A$  с использованием существующих систем обработки информации в сети интернет;

$\alpha_4(\phi, y)$  – перенос копии информационного носителя  $\phi$  с использованием средств доставки  $y$ ;

$\alpha_5(\mu, y)$  – отправка результата обработки информации заказчику  $\mu$  с использованием средств доставки  $y$ .

Отсюда, функция  $\psi$  опишется композицией элементарных функций:

$$\psi = \alpha_1(\mu, \omega) \square \alpha_2(b, g) \square \alpha_3(g, \Pi_A) \square \alpha_4(\phi, y) \square \alpha_5(\mu, y).$$

# Математическая постановка задачи поиска ИТИ сети интернет

## Первую группу задач

$$\psi_1 = (\mu(\omega), b(g), \ddot{I}_{A1}, \phi(y)), \mu \in M, \omega \in \Omega, b \in B, \Pi_{A1} \in \Pi_A, y \in Y, \phi \in \Phi, g \in G$$

$$\psi = \psi_1 = \alpha_1(\omega, \mu) \square \alpha_2(b, g) \square \alpha_3(g, \Pi_{A1}) \square \alpha_4(\phi, y) \square \alpha_5(\mu, y).$$

## Вторая группа задач

**Пертинентность** (*pertineo* — касаюсь, отношусь) — соответствие найденных информационно-поисковой системой документов информационным потребностям пользователя, независимо от того, как полно и как точно эта информационная потребность выражена в тексте информационного запроса.

$$\psi_2 = \alpha_1(\omega, \mu) \square \alpha_2(b, g) \square \alpha_3(g', \ddot{I}_{A2}) \square \alpha_4(\phi, y) \square \alpha_5(\mu, y) \square \alpha_6(\gamma, \lambda)$$

$\alpha_6(\gamma, \lambda)$  – разработка тематического запроса  $\gamma$  на аналитическую обработку информации  $\lambda$ .

# Информационная модель процесса обработки информации в сети интернет

Для представления процесса обработки информации в сети интернет в качестве информационного объекта  $z$  рассмотрим элементарный фрагмент информации информационного ресурса  $Z$ ,

Описать информационный объект (запись) можно четверкой:

$$z = (A_z, R_z, \Psi_z, L_z)$$

$A_z$  - алфавит (русский, латинский, ...);

$R_z$  - разделительные знаки (« » – пробел, :, -, ...);

$\Psi_z$  - словарь, являющийся языком в алфавите  $A$ :

$L_z$  - язык информационного объекта:

Произвольная совокупность записей ИР  $Z_1$ , также опишется четверкой:

$$Z_1 = (A_{Z_1}, R_{Z_1}, \Psi_{Z_1}, L_{Z_1}) = (\cup^{Z_1} A_z, \cup^{Z_1} R_z, \cup^{Z_1} \Psi_z, \cup^{Z_1} L_z) \in \Pi_Z$$

$\Pi_Z$  – полный набор всех возможных записей ИР характеризующего информацию находящуюся на данном ресурсе.

## Модель обработки информации в сети интернет

Обработка информации в сети интернет относится к классу итеративных задач.

$$\tilde{n} = t_{\text{э}} / \tilde{t}_l = t_{\text{э}} / \left( |\tilde{l}| \tilde{t}_{nn} \right)$$

$\tilde{t}_l$  - среднее время решения одной задачи;

$t_{\text{э}}$  - эфирное время обработки информации в РИС сети интернет;

$|\tilde{l}|$  - среднее число входящих в нее процедур обработки;

$\tilde{t}_{nn}$  - средняя продолжительность одной процедуры обработки.



## Модель обработки информации в сети интернет

Время одной процедуры обработки информации складывается из двух разнохарактерных по реализации составляющих:

$$\tilde{t}_{пп} = \tilde{t}_{пп}^a + \tilde{t}_{пп}^M$$

$\tilde{t}_{пп}^a$  - время затрачивается на аналитическую работу по построению очередной функции обработки информации,

$\tilde{t}_{пп}^M$  - время на ее решение в интернет, т.е. машинной составляющей, связанной с работой вычислительных и коммуникационных средств.

## Модель обработки информации в сети интернет

### Аналитическая составляющая лежит в основе:

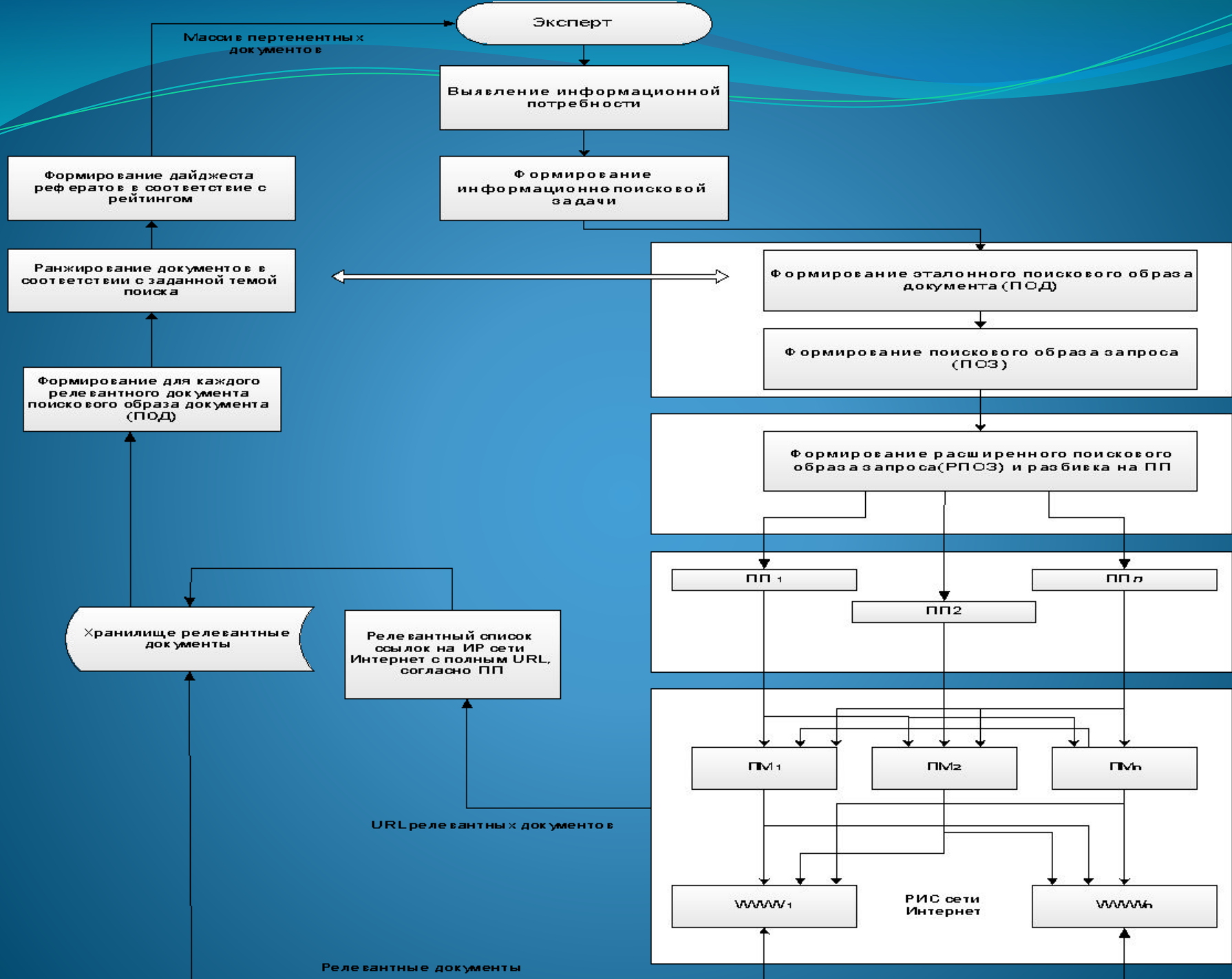
Формирования тезаурусов ПО, необходимых для составления поисковых предписаний (ПП) на обработку информации, и предусматривает использование имитационных моделей синонимии, дедукции и индукции.

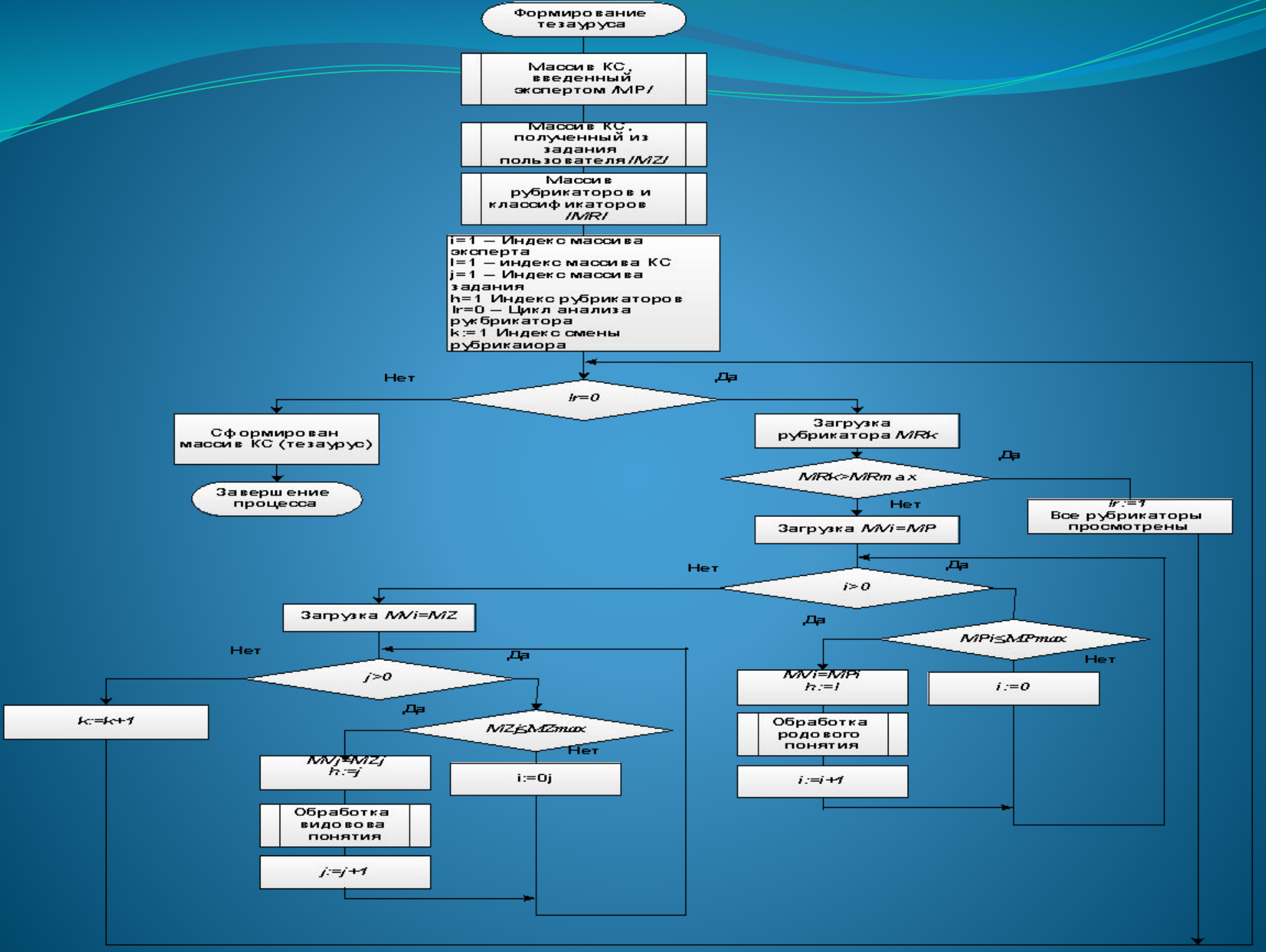
Синонимия используется для расширения ПО,

Дедукция – для формирования тезаурусов с использованием подхода от общего к частному,

Индукция – для формирования тезаурусов с использованием подхода от частного к общему.

Аналитическая составляющая базируется на использовании известных рубрикаторов (ГРНТИ, УДК, МПК и других). Тезаурусы формируются как тезаурусы КС и дескрипторов. Сформированные тезаурусы лежат в основе составления оптимальных ПП, отражающих в своей сути искомый поисковый образ документа (ПОД) в РИС сети интернет.





1. Анализ и краткое описание предметной области.
2. Выявление(определение) информационной потребности.
3. Формирование информационно-поисковых задач.
4. Определение поисковой стратегии:
  - Поиска структурированной информации
    - Определение предметной области поиска информации (ГРНТИ, УДК, МПК и т.п.)
    - Составление списка ключевых слов.
    - Формирование дескрипторов
    - Формирование поисковых предписаний
  - Поиска неструктурированной информации
    - Определение предметной области поиска информации (ГРНТИ, УДК, МПК и т.п.)
    - Составление списка ключевых слов.
    - Формирование дескрипторов (Эталонных)
    - Формирование словоформ (Словарь Зализняка)
    - Формирование синонимии (Синонимов)
    - Расширение предметной области поиска информации ((ГРНТИ, УДК, МПК и т.п., по методу дедукции и индукции))
    - Формирование дескрипторов с учетом синонимов и расширения ПО
    - Формирование поисковых предписаний



- Ключевые слова
- Дескрипторы
- Ключевые слова и словоформы
- Дескрипторы и синонимы
- Дескрипторы и синонимы с учетом расширенной предметной области
- Все перечисленные выше комбинации вместе.

5. Выбор инструментов и механизмов поиска информации:

- Структурированной информации
- Неструктурированной информации

6. Осуществление поиска информации.

- Комбинация поискового предписания  $n, n+1, \dots$
- Результат поиска (Релевантный, только ссылки, не менее 200 просмотренных ссылок)
- Результат поиска (Пертинентный, полный текст)
- Процент результативности (Отношение релевантных документов к пертинентным)

7. Выводы по информационному поиску