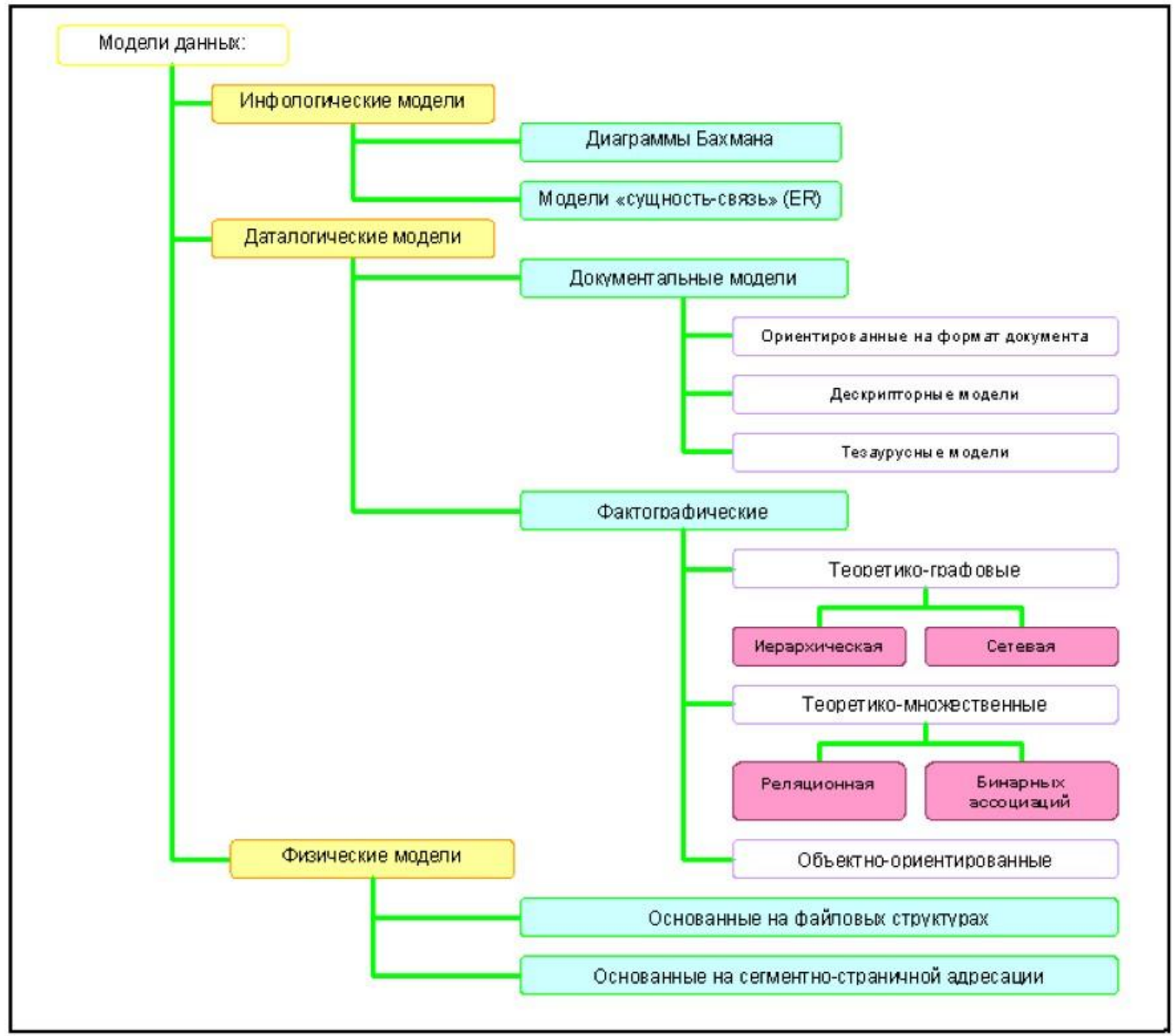
Документальные информационные системы


Модели данных:
Инфологические модели
Диаграммы Бахмана
Модели «сущность-связь» (ER)
Даталогические модели
Документальные модели
Ориентированные на формат документа
Дескрипторные модели
Тезаурусные модели
Фактографические
Теоретико-графовые
Иерархическая
Сетевая
Теоретико-множественные
Реляционная
Бинарных ассоциаций
Объектно-ориентированные
Физические модели
Основанные на файловых структурах
Основанные на сегментно-страничной адресации

Документальная база данных — база, в которой каждый элемент представлен только в объёме поискового образа адресата. Такой элемент содержит совокупность признаков, характеризующих адресата (полный почтовый адрес, банковские реквизиты, фамилии-имена-отчества нужных лиц, профиль деятельности и т.д.).

В *документальных БД* единицей хранения является документ и пользователю выдается ссылка на документ или сам документ. Документальные БД организуются без хранения и с хранением документа на машинных носителях. К первому типу относятся *библиографические*, *реферативные* и *БД-указатели*, отсылающие к источнику информации. Системы, хранящие полный текст документа, называются *полнотекстовыми.* Их разновидностью являются *БД форм документов*, в которых документ ищется для использования его в качестве шаблона.

К *лексикографическим БД* относятся различные словари (классификаторы, многоязычные словари, словари основ слов и т. п.).

Документальные модели данных соответствуют представлению о слабоструктурированной информации, ориентированной в основном на свободные форматы документов, текстов на естественном языке.
Модели, ориентированные на формат документов, связаны прежде всего со стандартным общим языком разметки — SGML (Standart Generalised Markup Language), который был утвержден ISO в качестве стандарта еще в 80-х годах. Этот язык предназначен для создания других языков разметки, он определяет допустимый набор тегов (ссылок), их атрибуты и внутреннюю структуру документа. Контроль за правильностью использования тегов осуществляется при помощи специального набора правил, которые используются программой клиента при разборе документа. Для каждого класса документов определяется свой набор правил, описывающих грамматику соответствующего языка разметки. Гораздо более простой и удобный, чем SGML, язык HTML (HyperText Markup Language – язык разметки гипертекста) позволяет определять оформление элементов документа и имеет некий ограниченный набор инструкций — тегов, при помощи которых осуществляется процесс разметки. Инструкции HTML в первую очередь предназначены для управления процессом вывода содержимого документа на экране программы-клиента и определяют этим самым способ представления документа, но не его структуру. В качестве элемента гипертекстовой базы данных, описываемой HTML, используется текстовый файл, который может легко передаваться по сети с использованием протокола HTTP. В настоящее время все большую популярность приобретает язык XML (eXtensible Markup Language – расширяемый язык разметки), позволяющий описывать документы произвольной структуры и содержания.
Тезаурусные модели основаны на принципе организации словарей. Они содержат определенные языковые конструкции и принципы их взаимодействия в заданной грамматике. Эти модели эффективно используются в системах-переводчиках, особенно многоязыковых. Принцип хранения информации в этих системах и подчиняется тезаурусным моделям.

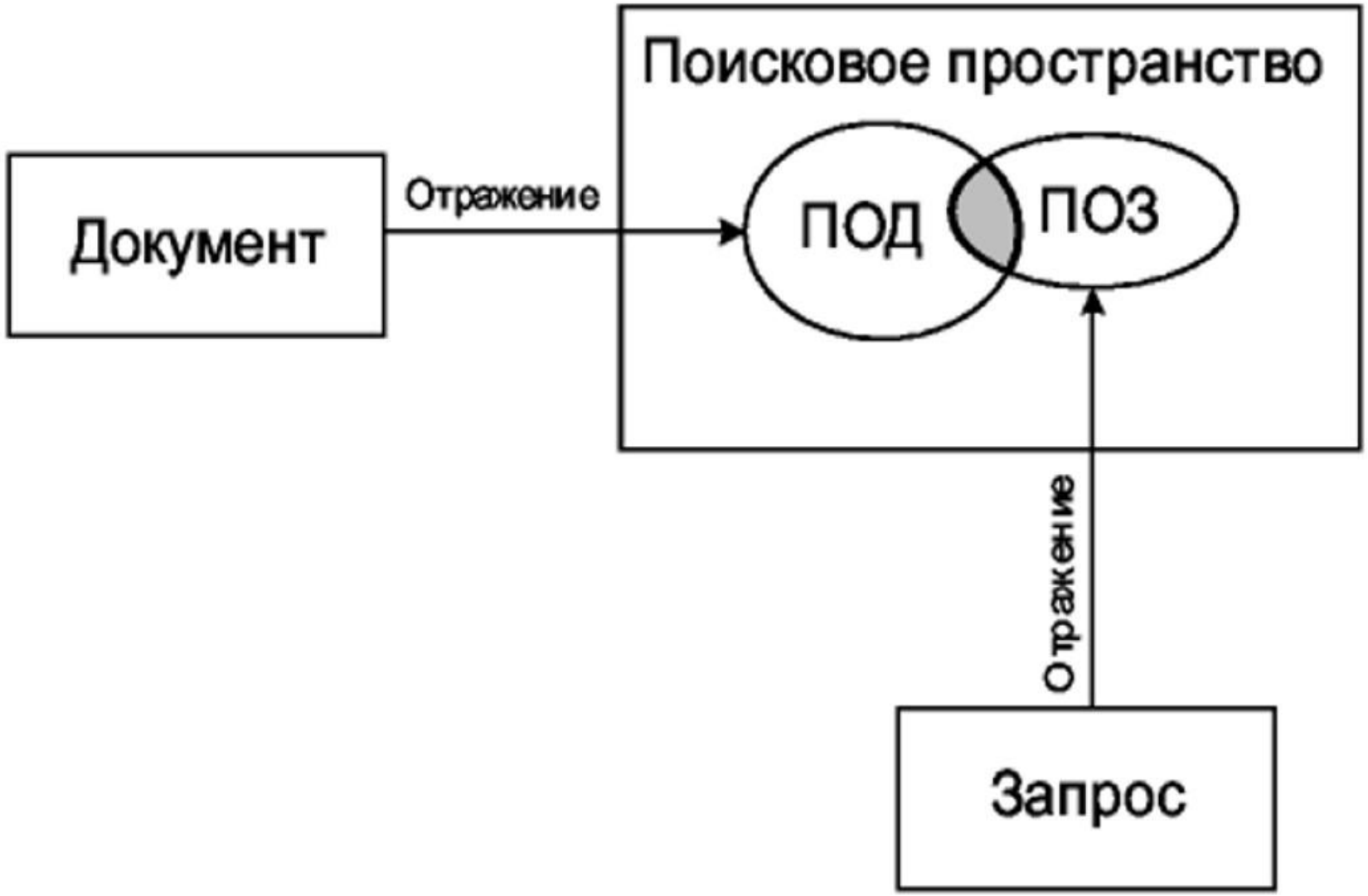
Дескриптпорные модели — самые простые из документальных моделей, они широко использовались на ранних стадиях использования документальных баз данных. В этих моделях каждому документу соответствовал дескриптор — описатель. Этот дескриптор имеет жесткую структуру и описывает документ в соответствии с теми характеристиками, которые требуются для работы с документами в разрабатываемой документальной базе данных. Например, для БД, содержащей описание патентов, дескриптор содержит название области, к которой относился патент, номер патента, дату выдачи патента и еще ряд ключевых параметров, которые заполнялись для каждого патента. Обработка информации в таких базах данных ведется исключительно по дескрипторам, то есть по тем параметрам, которые характеризуют патент, а не по самому тексту патента.

Теоретико-графовые модели отражают совокупность объектов реального мира в виде графа взаимосвязанных информационных объектов. Математической основой таких моделей является теория графов. Реляционная модель будет подробно рассмотрена далее.

Модели, ориентированные на формат документов, связаны прежде всего со стандартным общим языком разметки — SGML (Standart Generalised Markup Language), который был утвержден ISO в качестве стандарта еще в 80-х годах. Этот язык предназначен для создания других языков разметки, он определяет допустимый набор тегов (ссылок), их атрибуты и внутреннюю структуру документа. Контроль за правильностью использования тегов осуществляется при помощи специального набора правил, которые используются программой клиента при разборе документа. Для каждого класса документов определяется свой набор правил, описывающих грамматику соответствующего языка разметки. Гораздо более простой и удобный, чем SGML, язык HTML (HyperText Markup Language – язык разметки гипертекста) позволяет определять оформление элементов документа и имеет некий ограниченный набор инструкций — тегов, при помощи которых осуществляется процесс разметки. Инструкции HTML в первую очередь предназначены для управления процессом вывода содержимого документа на экране программы-клиента и определяют этим самым способ представления документа, но не его структуру. В качестве элемента гипертекстовой базы данных, описываемой HTML, используется текстовый файл, который может легко передаваться по сети с использованием протокола HTTP. В настоящее время все большую популярность приобретает язык XML (eXtensible Markup Language – расширяемый язык разметки), позволяющий описывать документы произвольной структуры и содержания.

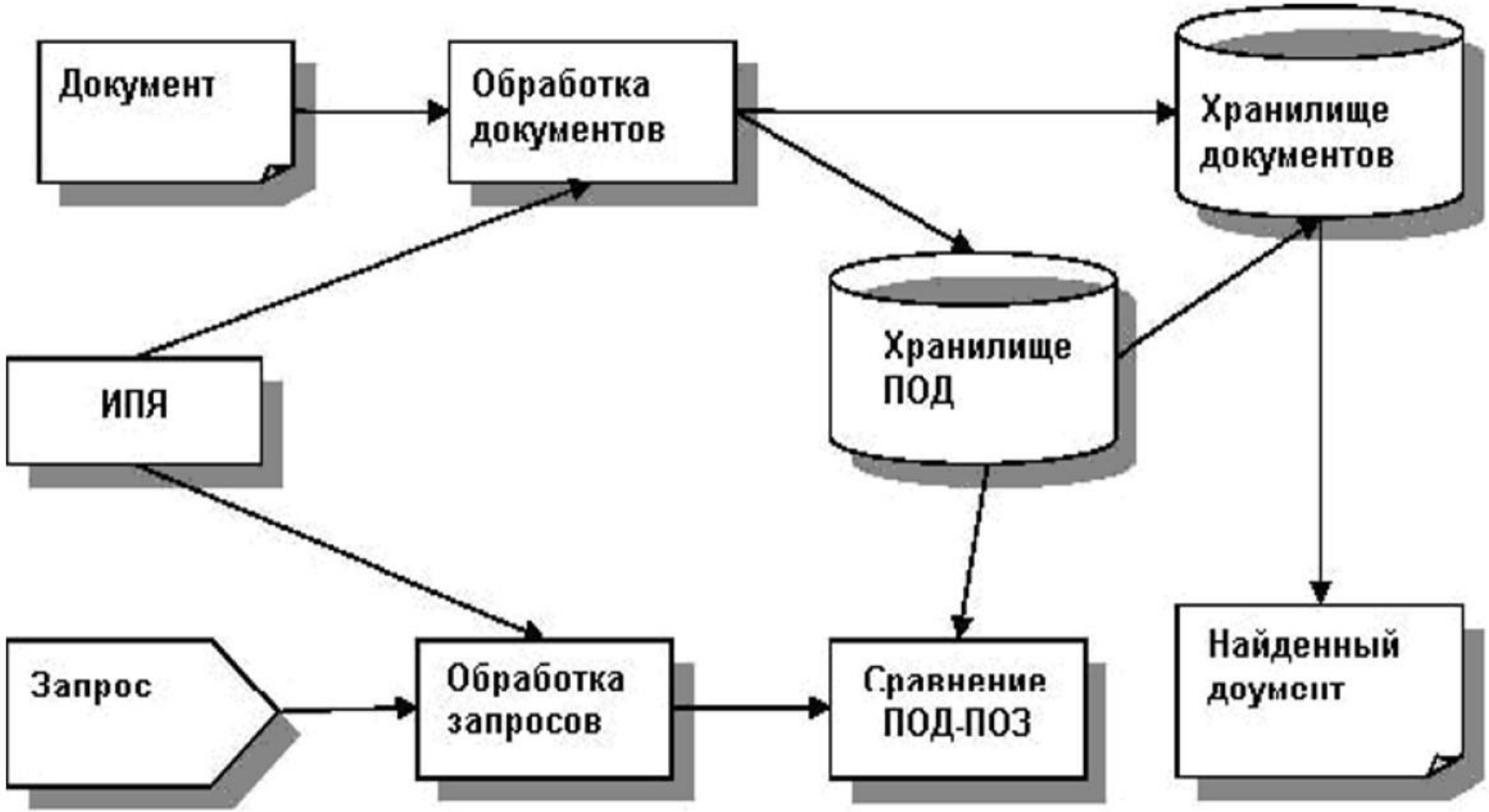
Тезаурусные модели основаны на принципе организации словарей. Они содержат определенные языковые конструкции и принципы их взаимодействия в заданной грамматике. Эти модели эффективно используются в системах-переводчиках, особенно многоязыковых. Принцип хранения информации в этих системах и подчиняется тезаурусным моделям.

Дескриптпорные модели — самые простые из документальных моделей, они широко использовались на ранних стадиях использования документальных баз данных. В этих моделях каждому документу соответствовал дескриптор — описатель. Этот дескриптор имеет жесткую структуру и описывает документ в соответствии с теми характеристиками, которые требуются для работы с документами в разрабатываемой документальной базе данных. Например, для БД, содержащей описание патентов, дескриптор содержит название области, к которой относился патент, номер патента, дату выдачи патента и еще ряд ключевых параметров, которые заполнялись для каждого патента.

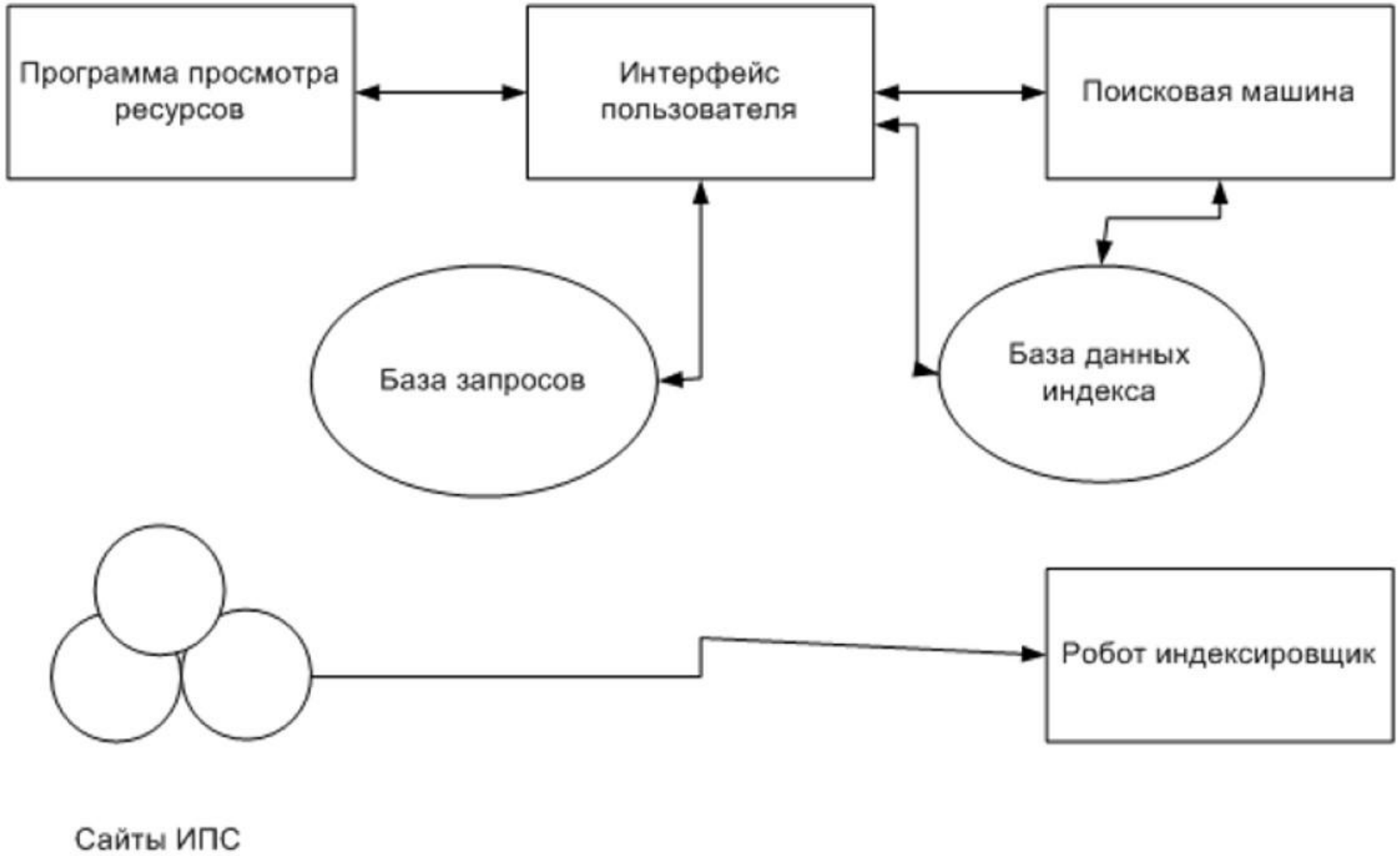


Общий принцип функционирования документальных ИПС на основе индексирования

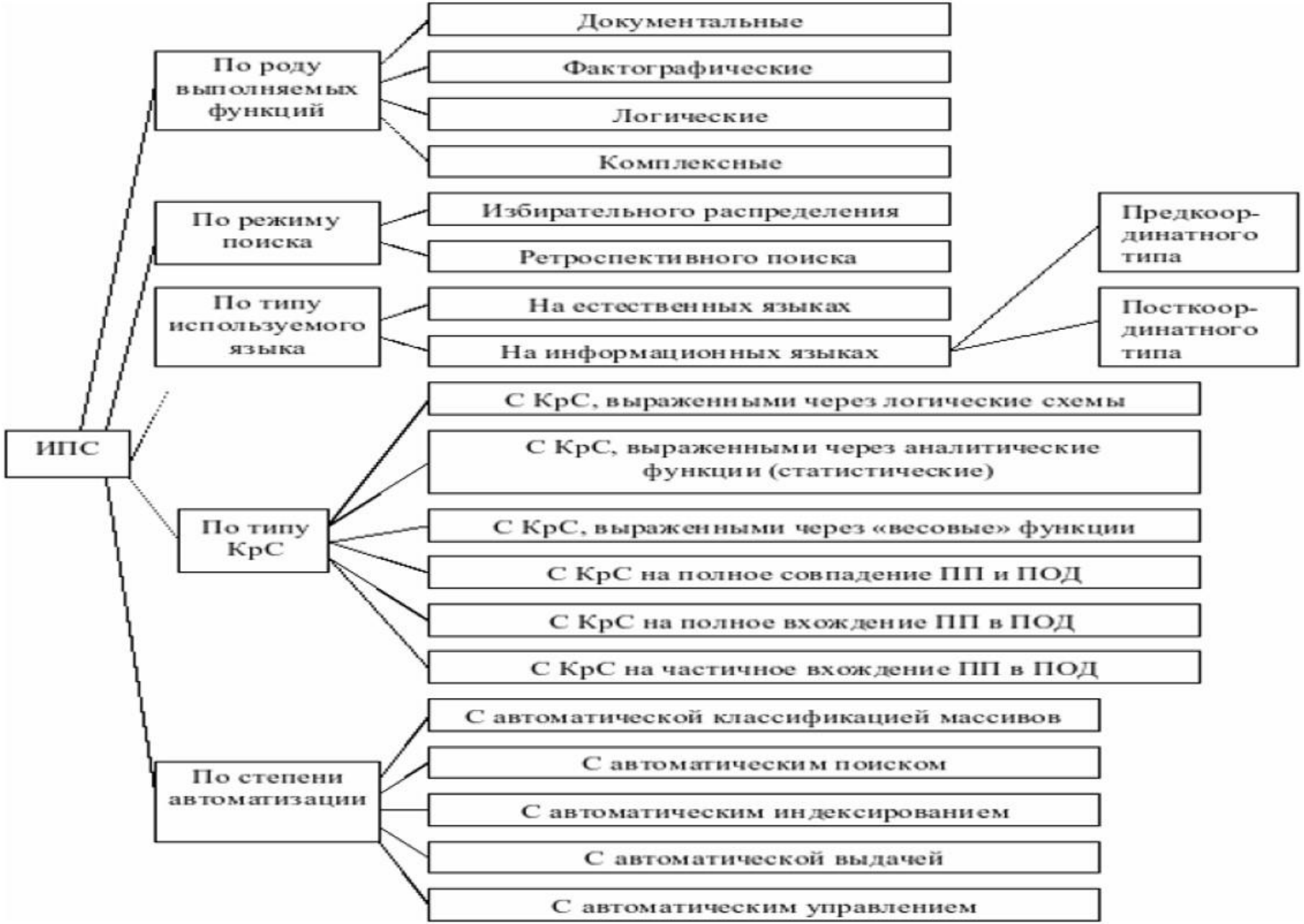

Поисковые образы и адреса
Активное запоми-
нающее устройство
Индексирование
запросов
Запросы
Документы
Индексирование
документов
Решающее
устройство
Документы
с адресами
Адреса
Пассивное запоми-
нающее устройство
Документы



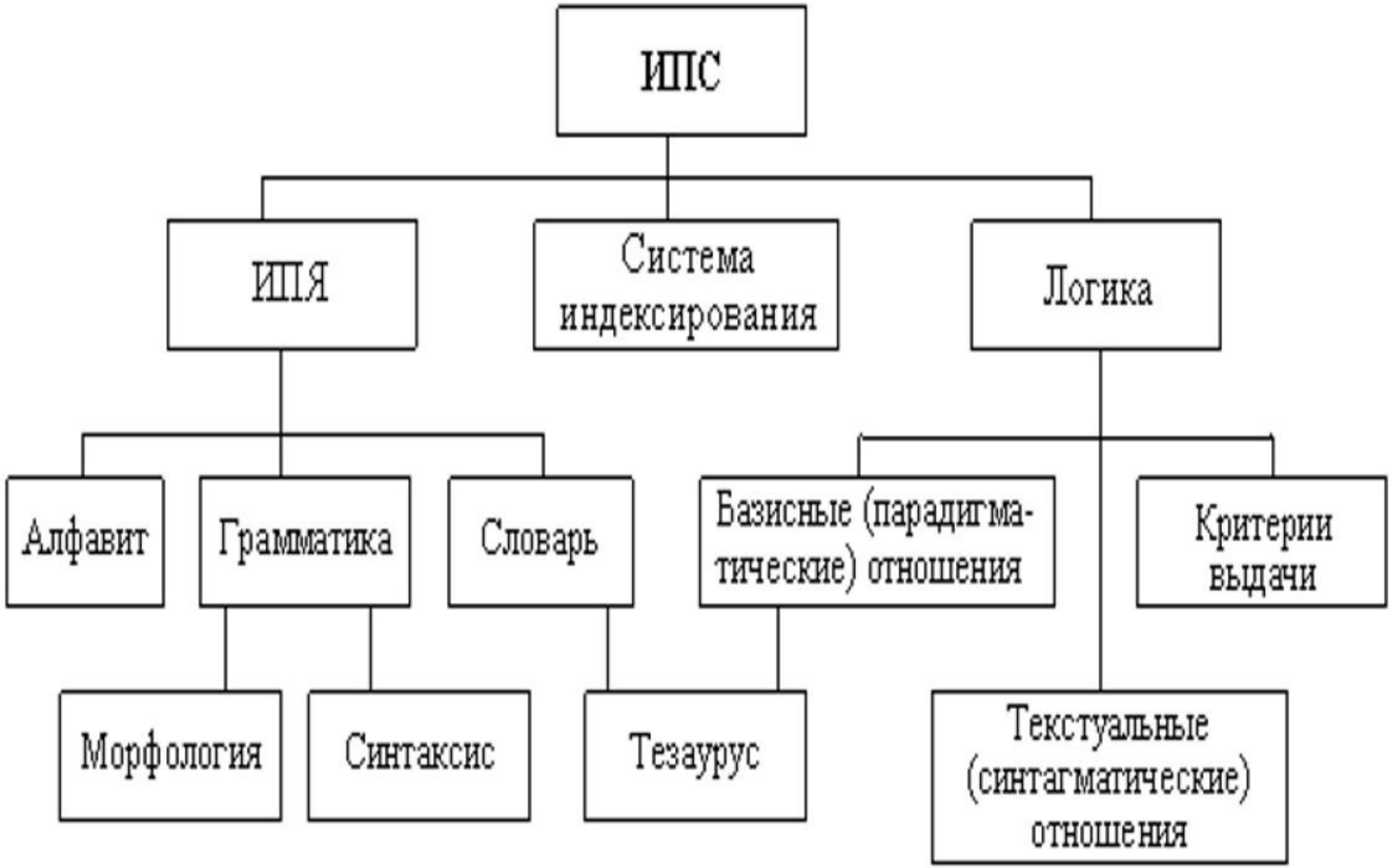
Структура информационных потоков при поиске документов в информационно-поисковой системе



Структурная система ИПС для Интернета



Классификация информационно-поисковых систем по роду выполняемых функций.


ИПС
ИПЯ
Система индексирования
Логика
Алфавит
Грамматика
Словарь
Базисные (парадигма-тические) отношения
Критерии выдачи
Морфология
Синтаксис
Тезаурус
Текстуальные (синтагматические) отношения

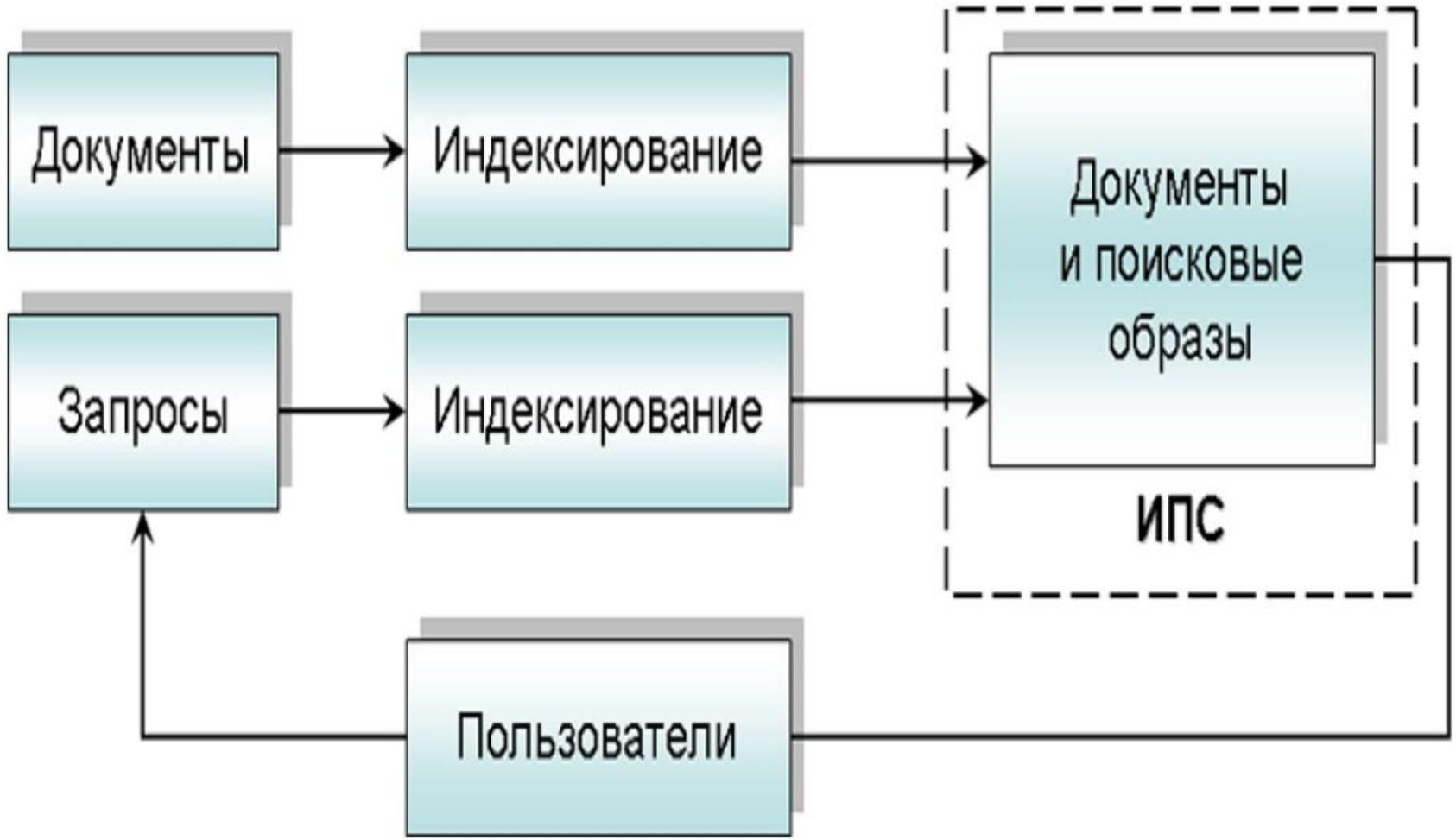


Схема информационно-поисковой системы