ПОСТРОЕНИЕ ПОИСКОВОГО ОБРАЗА ДОКУМЕНТА НА ОСНОВЕ ЕГО АВТОМАТИЧЕСКОГО АНАЛИЗА

В соответствии с решаемой задачей всем частям речи были приписаны ЛГК для определения принадлежности слова к конкретной части речи, получения канонической формы и полной словоизменительной парадигмы.

Алгоритм построения ПОД для каждого *j*-го текстового документа, $j=\overline{1,l}$, состоит в построении списка $D^{(j)}$ информативных слов:

Шаг 1. Из текстового документа выбрать только информативные слова. Выбор этих слов выполняется по их ЛГК с фильтрацией неинформативных слов по кодам и по стоп-словарю.

Шаг 2. Для каждого *i*-го слова $w_i^{(j)}$ из списка $D^{(j)}$ выполнить процедуру канонизации.

Шаг 3. для всех информативных слов $w_i^{(j)}$ *j*-го документа, $i=\overline{1,n^{(j)}}$, вычислить их абсолютные частоты $\nu_i^{(j)}$.

Шаг 4. вычислить вес $p_i^{(j)}$ каждого слова $w_i^{(j)}$:

$$p_i^{(j)}=\frac{\nu_i^{(j)}}{\nu_{\max}^{(j)}}$$

Релевантным ЕЯ-запросу Q считается каждый η–й документ из БД, для которого

$$D^{(\eta)}\cap Q\neq \varnothing$$

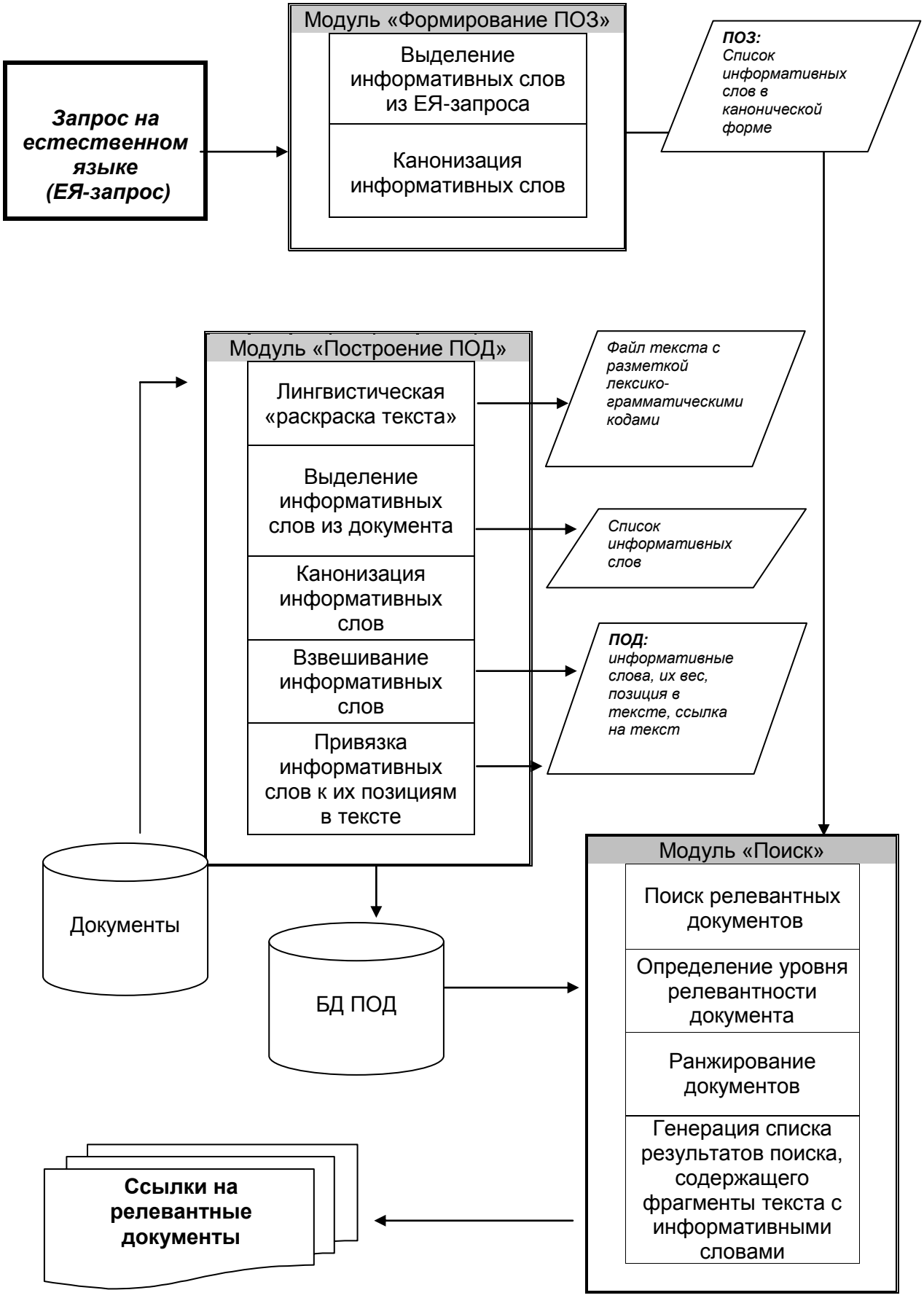
Обозначим

$$D^{(\eta)}\cap Q=R^{(\eta)},\quad \left|R^{(\eta)}\right|=K^{(\eta)},\quad \eta=\overline{1,q}$$

Уровень r_η релевантности ЕЯ-запросу каждого η–го текстового документа, $\eta=\overline{1,q}$ определяем по формуле:

$$r_\eta=\sum_s p_s^{(\eta)},\quad s=\overline{1,k^{(\eta)}}$$

Все полученные релевантные документы ранжируем в соответствии с убыванием уровня релевантности r_η для каждого η–ого документа.


Модуль «Формирование ПОЗ»
Выделение информативных слов из ЕЯ-запроса
Канонизация информативных слов
Запрос на естественном языке (ЕЯ-запрос)
ПОЗ: Список информативных слов в канонической форме
Модуль «Построение ПОД»
Лингвистическая «раскраска текста»
Выделение информативных слов из документа
Канонизация информативных слов
Взвешивание информативных слов
Привязка информативных слов к их позициям в тексте
Файл текста с разметкой лексико-грамматическими кодами
Список информативных слов
ПОД: информативные слова, их вес, позиция в тексте, ссылка на текст
Документы
БД ПОД
Модуль «Поиск»
Поиск релевантных документов
Определение уровня релевантности документа
Ранжирование документов
Генерация списка результатов поиска, содержащего фрагменты текста с информативными словами
Ссылки на релевантные документы