ТРУСОВ Александр Владимирович -
кандидат технических наук, доцент, директор Пермского ЦНТИ - филиала ФГУ «Российское энергетическое агентство» Минэнерго России
Адрес: 614600, г. Пермь, ул. Попова, 9
e-mail: tav@permcnti.ru

ТРУСОВ Владимир Александрович -
аспирант ГОУ ВПО «Пермский государственный технический университет»
Адрес: 614990, г. Пермь, Комсомольский проспект, 20
e-mail: tva@permcnti.ru

ПОДХОДЫ К ФОРМИРОВАНИЮ СМЫСЛОВОГО ПОИСКА ИНФОРМАЦИИ В РАСПРЕДЕЛЕННЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ СЕТИ ИНТЕРНЕТ

Интерес к вопросу о поиске информации в интернете не ослабевает на протяжении всего времени существования сети. Поиск может вестись как пользователем-любителем, так и профессионалом. При проведении поиска информации, удовлетворяющего информационным потребностям пользователя, необходимо знать, от чего зависит успешный поиск и какие проблемы возникают при работе с информацией.

Обобщенный алгоритм поиска информации в распределенных информационных системах (РИС) сети Интернет представлен на **рис. 1**.

Поиск информации в РИС сети Интернет может быть проведен несколькими методами, значительно различающимися как по эффективности и качеству поиска, так и по типу извлекаемой информации [1].

Независимо от того, кто делает запрос поисковой системе, пользователь-профессионал или просто любитель, он ждет от системы предоставления документов, максимально соответствующих его запросу (поисковому образу), то есть релевантных [2]. Понятие «релевантность» возникло с развитием теории информационного поиска. Релевантность определяется базисными понятиями: информационный поиск, информационно-поисковые системы, информационный запрос, текст, выдача.

Однако в большинстве случаев релевантных запросу записей в лучшем случае сотни, а в худшем случае тысячи, из которых лишь несколько удовлетворяют смыслу запроса. Отсюда и возникает понятие «пертенентной» информации, т.е. информации, удовлетворяющей смыслу запроса.

К поисковой системе предъявляются три основных требования:

- контроль полноты охвата ресурсов;
- контроль достоверности информации, полученной из сети;
- высокая скорость проведения поиска.

Контроль полноты охвата ресурсов

При проведении поиска информации в сети Интернет по какому-либо вопросу используются различные типы ресурсов. Знание всех основных существующих на сегодняшний день типов ресурсов сети, понимание технической и тематической специфики их информационного наполнения и особенностей доступа становится необходимым условием успешного планирования и проведения поисковых работ.

Контроль достоверности информации

Контроль может производиться разными средствами. Традиционными способами проверки являются:

- локализация источников информации, альтернативных поисковым данным;
- сверка фактического материала, установление частоты его использования другими источниками;
- выяснение статуса документа и рейтинга узла, на котором он находится, средствами поисковых систем;
- получение информации о компетентности и статусе автора материала с помощью специальных поисковых сервисов;
- анализ отдельных элементов организации узла с целью оценки квалификации специалистов, его поддерживающих, и другие.

Скорость проведения поиска в сети

Скорость проведения поиска в сети зависит от двух факторов:

- от грамотного планирования поисковой процедуры;
- от навыков работы с ресурсом выбранного типа.

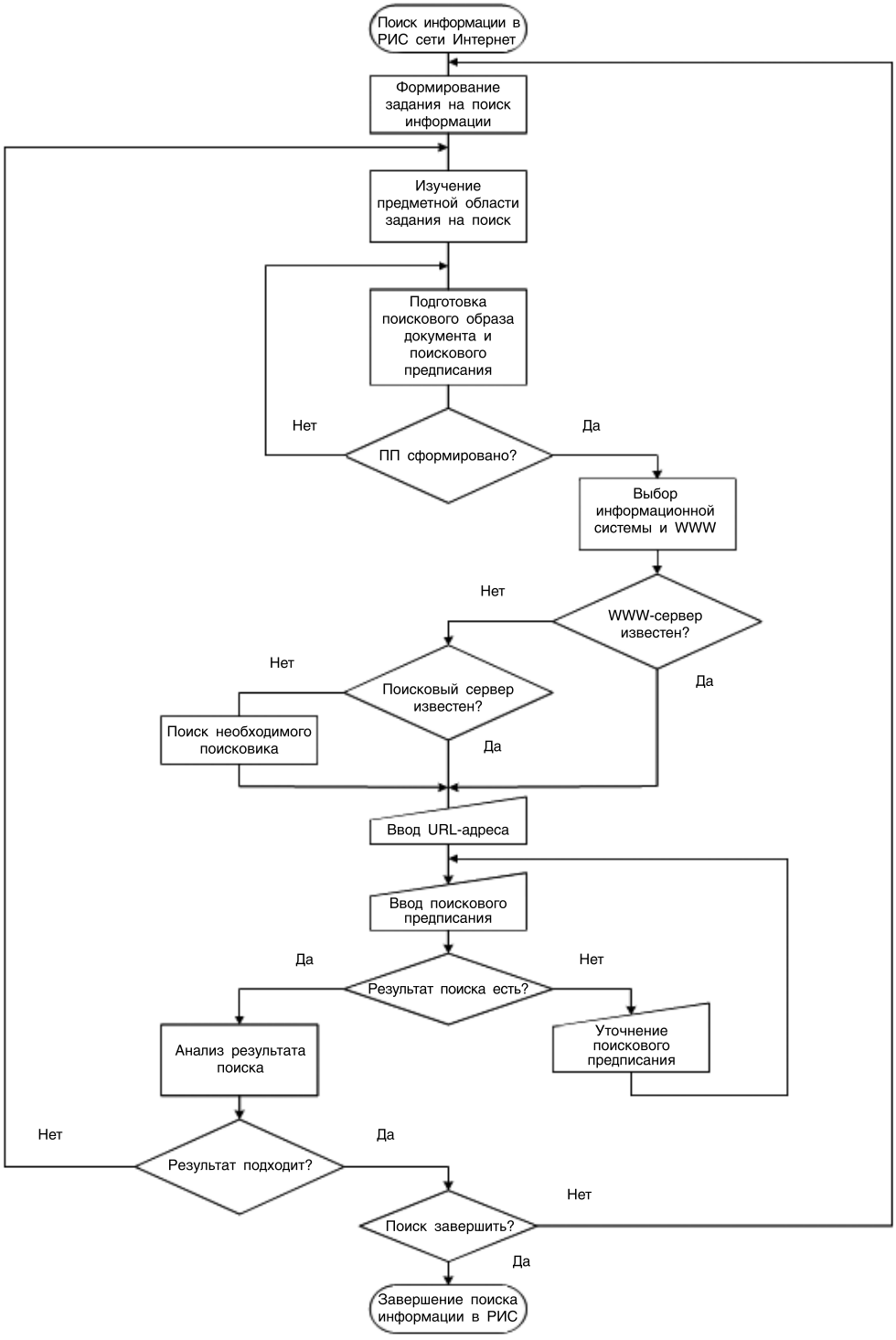

Рис. 1. *Алгоритм поиска информации в РИС сети Интернет*

Под составлением плана поисковых работ понимается выбор поисковых сервисов и инструментов, отвечающих специфике задачи и последовательности их применения в зависимости от ожидаемой результативности. После получения доступа к соответствующему ресурсу важно суметь быстро разобраться в его структуре и способах навигации. Моторика выполнения действий, умелое совмещение поисковых средств и возможностей обработки информации локальной клиентской программы и сервера для поисковика являются необходимыми навыками.

Поиск информации с использованием механизма синонимии

Поиск информации с использованием синонимии включает:

1) анализ задания на поиск информации, заданной предметной области (*ПО*), ключевых слов (*КС*) и дескрипторов (*Д*);

2) поиск информации с использованием механизма синонимии.

Одним из важнейших элементов влияющего на результаты поиска информации является тезаурус ключе-

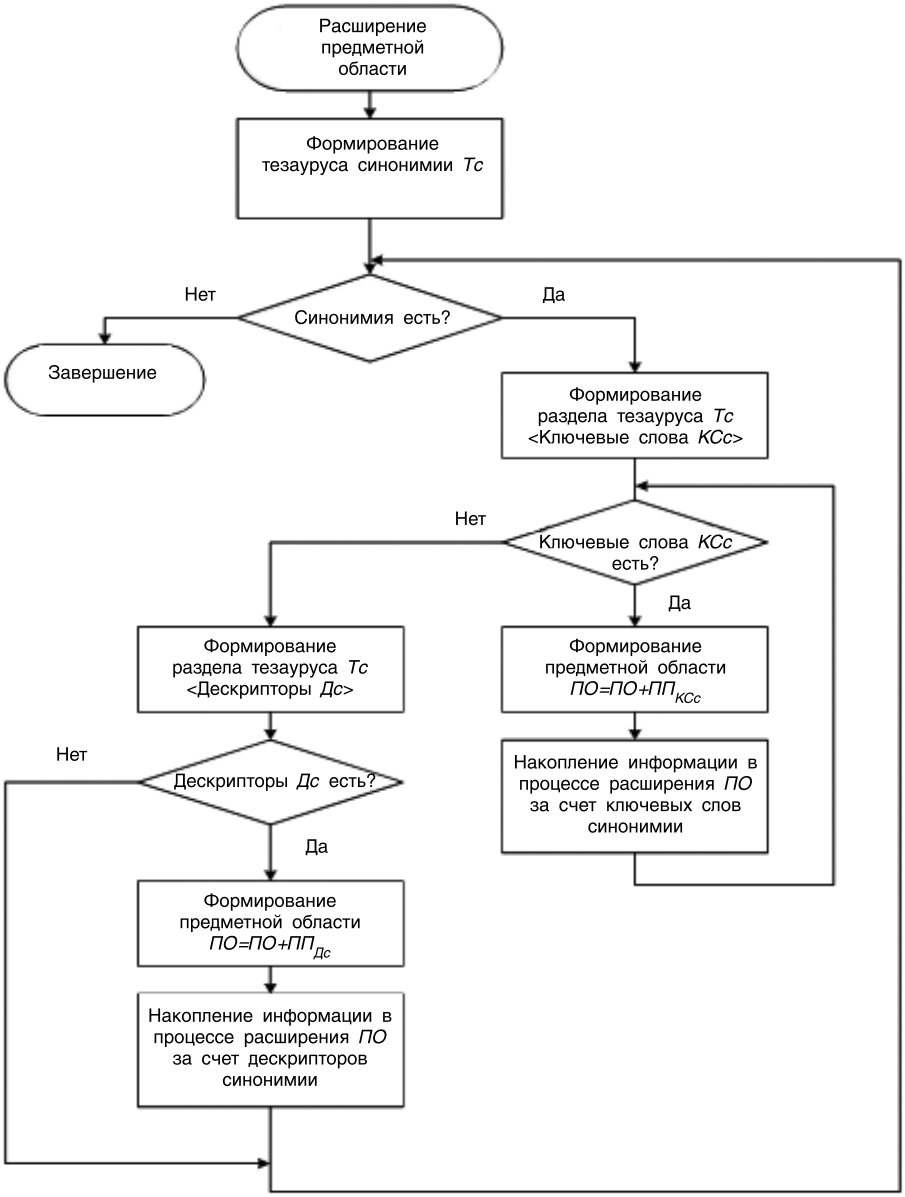

Рис. 2. *Алгоритм расширения предметной области с использованием синонимии*

вых слов, который включает в себя расширение предметной области за счет синонимии и формирование на этой основе тезауруса синонимии (*Тс*).

Расширение предметной области за счет синонимии (**рис. 2**) включает:

1) формирование тезауруса синонимии;

2) разделы тезауруса по ключевым словам или дескрипторам;

3) формирование предметной области;

4) накопление информации в процессе расширения предметной области за счет синонимии.

Формирование тезауруса синонимии (**рис. 3**) включает:

1) анализ предметной области, ключевых слов и дескрипторов;

2) определение синонимии в заданной предметной области;

3) формирование тезауруса синонимии заданной предметной области;

4) формирование разделов ключевых слов или дескрипторов тезауруса синонимии заданной предметной области.

Поиск информации в распределенных информационных ресурсах сети Интернет реализуется по принципу «от простого к сложному», т.е. предполагает постепенное погружение специалиста-эксперта в процесс решения информационно-поисковых задач, связанных с использованием синонимик все более сложного характера.

Оценка свойств распределенных информационных систем сети Интернет характеризуется оперативностью, динамичностью и информативностью.

Оперативность характеризуется быстротой доступа специалиста-эксперта к информационным ресурсам РИС сети Интернет:

$$\tilde{t}^{M}_{nn} = \tilde{t}^{M}_{-} + \tilde{t}^{M}_{\sim}, \quad (1)$$

$$\tilde{t}^{a}_{nn} = \tilde{t}^{a}_{\{z\}} + \tilde{t}^{a}_{\varphi}, \quad (2)$$

где $\tilde{t}^{M}_{-}$ - среднее время выполнения ПП в РИС сети Интернет без учета потерь в сети ($\tilde{t}^{M}_{-}$);

$\tilde{t}^{a}_{\{z\}}$ - среднее время выборочного анализа записей $\{z\}$ из текущего состояния РИС;

$\tilde{t}^{a}_{\varphi}$ - среднее время составления ПП (аналитическая составляющая).

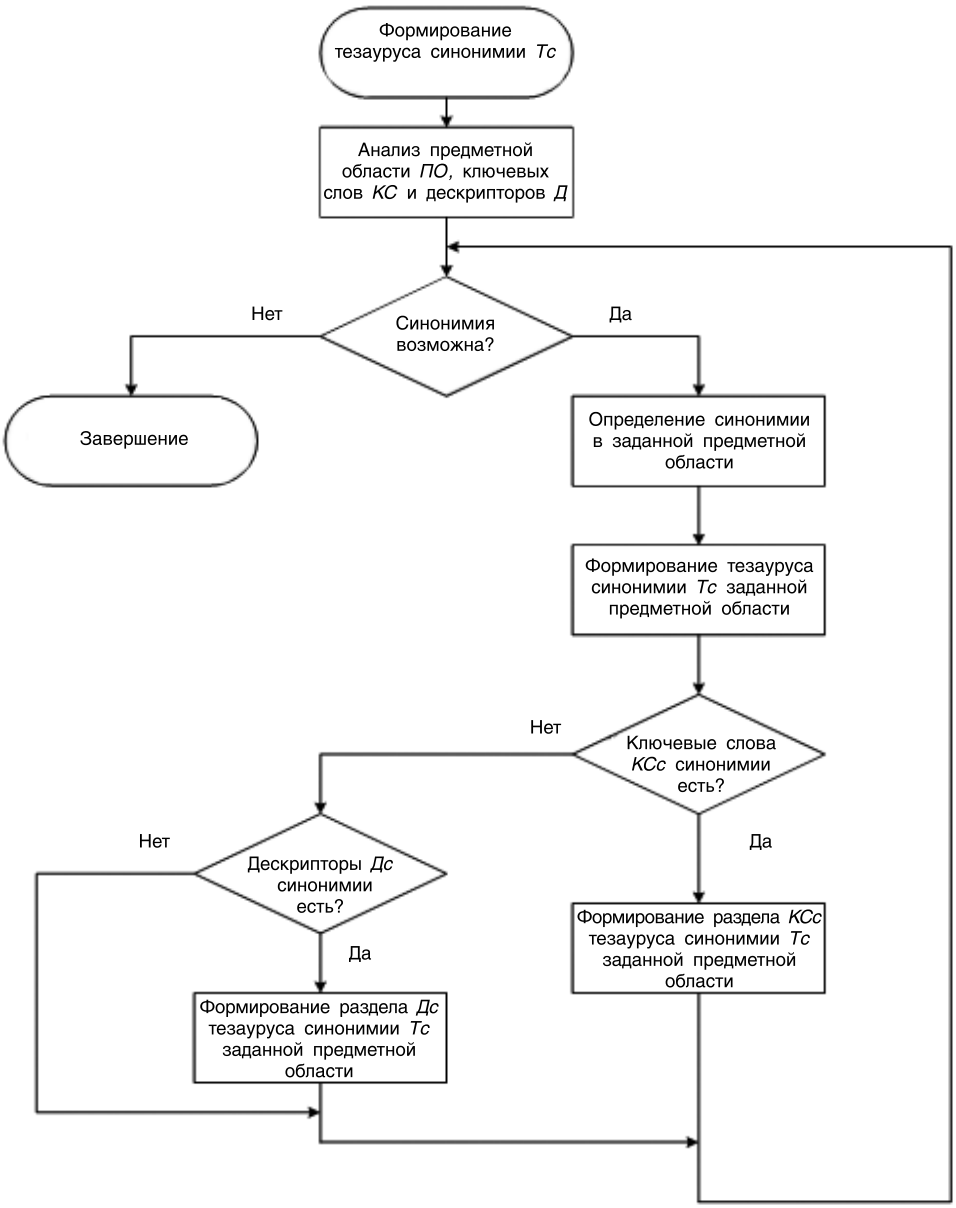

Рис. 3. *Алгоритм формирования тезауруса синонимии*

Динамичность характеризуется мощностью наборов синонимик $\{n_c^{max}\}$, антиномий $\{n_a^{max}\}$; возможностью варьирования в широком диапазоне показателями точности θ и полноты π поиска, правда, при большой проблематичности их даже приближенного оценивания и, тем более, документирования.

Информативность характеризуется объемами предоставляемой специалисту-эксперту из РИС информации. Только теоретически они определяются возможностями самих РИС сети Интернет, а на практике также ограничиваются предоставляемым временем работы в сети Интернет.

На основании проведенных исследований можно предложить следующий алгоритм поиска информации в распределенных информационных ресурсах сети Интернет.

Поиск информации в РИС сети Интернет включает:

1) максимальное покрытие решения задачи $\rho^{\bar{z}}$ поисковой процедурой, составленной из известных оператору ключевых слов тезауруса предметной области $T^{O}_{ОП}$ ($T^{З}_{ПО}$)

$$\varphi: \vee^{n_{max}} P_n(t_n = t_n^z) : Z \rightarrow \rho^z_{hi},\ h=1\ ; \qquad (3)$$

2) исследование $\{z\}$ ρ^z_{hi} на предмет выявления антиномий n_a, синонимик n_c и дескрипторов компонент связности $d_{св}, d_д$;

3) антиномии разрешаются поисковыми процедурами вида:

$$\varphi := t_a \& \bar{d} \mid t_a \& \& \bar{d}, \qquad (4)$$

где d - дескриптор нежелательной компоненты связности;

4) компоненты связности $d_{св}$, не принадлежащие решению $\rho^{\bar{z}}$, удаляются ПП вида:

$$\varphi := d_{св} \& \bar{d}_д \mid d_{св} \& \& \bar{d}_д; \qquad (5)$$

5) синонимики n_c используются для расширения «покрытия» решения задачи $\rho^{\bar{z}}$ процедурой вида (3) при

$$n_{max} := n_{max} + n_c \qquad (6)$$

На основе предложенного алгоритма поиска информации разработана модель, дающая возможность перейти к автоматизированному смысловому поиску информации в распределенных информационных системах сети Интернет (**рис. 4**).

Ценность любой информационной сети определяется ее информационными ресурсами, то есть знаниями и данными, которые сеть поставляет пользователю. Эти ресурсы должны как можно шире охватывать те области, в которых работают пользователи сети.

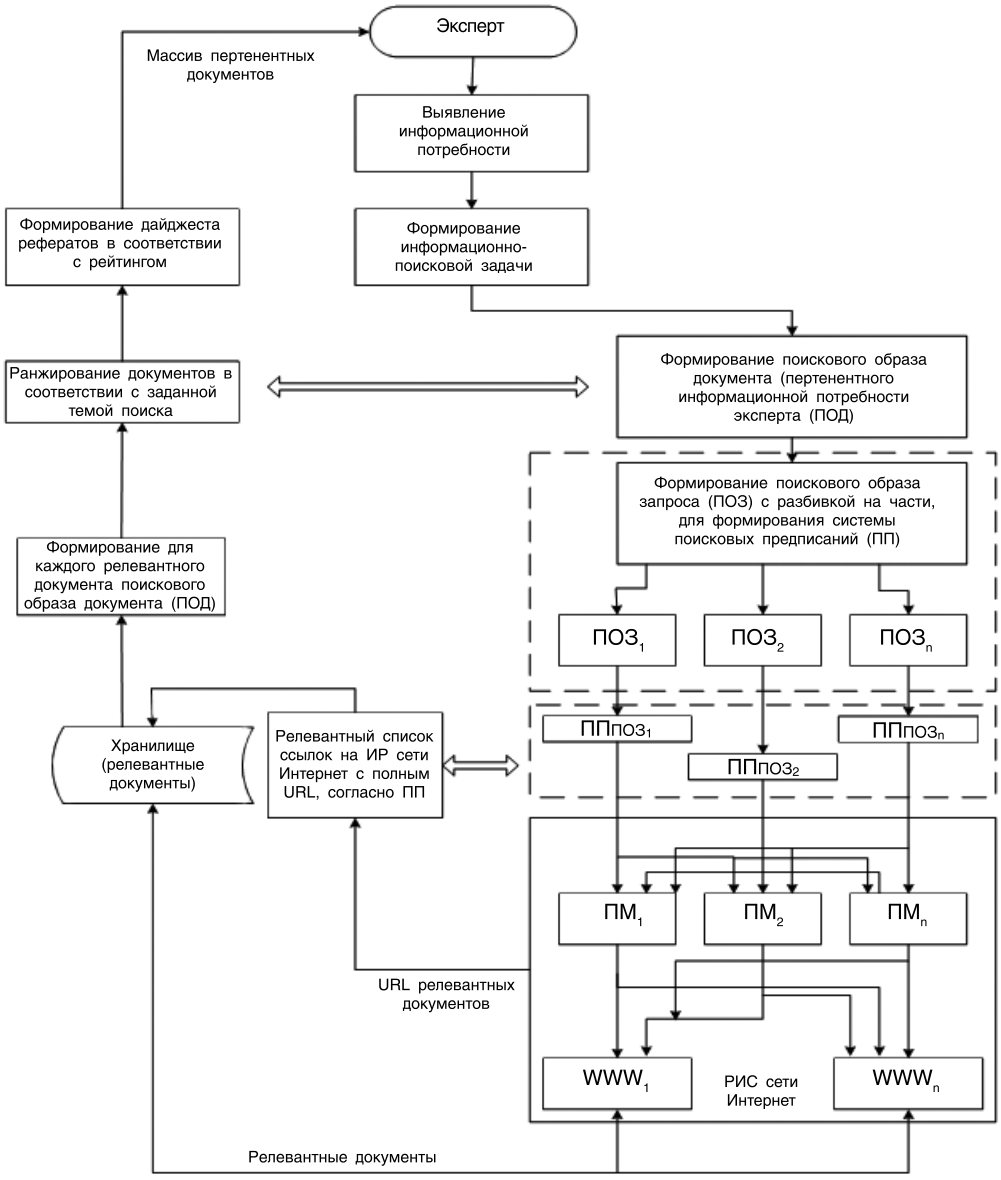

Рис. 4. *Модель смыслового поиска информации в распределенных информационных системах сети Интернет*

Обобщенный алгоритм реализации модели смыслового поиска информации в распределенных информационных системах сети Интернет выглядит следующим образом:

1. «Образец документа» (смысловая задача), представляющий собой шаблон поиска, вводится экспертом вручную.
2. Из документа выделяется тема запроса, и определяются поисковые предписания.
3. Расширяется тема запроса за счет синонимии и ассоциативных запросов.
4. Формируется поисковый образ запроса (ПОЗ) на основе частотного словаря с разбивкой его на отдельные поисковые предписания.
5. Производится первоначальный поиск ссылок на релевантные документы в существующих поисковых Интернет-машинах, общий результат помещается в хранилище данных.
6. Осуществляется закачка найденных документов в хранилище данных.
7. Для каждого документа в хранилище данных формируется поисковый образ документа (ПОД).
8. Производится ранжирование документов в соответствии с заданной темой (п. 2). Чем концентрированнее смысл в теле документа, тем он имеет более высокий рейтинг.
9. Производится реферирование найденных документов и осуществляется передача рефератов для ознакомления и анализа эксперту в соответствии с рейтингом.

Предложенный подход к организации смыслового поиска информации в распределенных информационных системах сети Интернет позволяет качественно улучшить результаты поисковых запросов к поисковым машинам сети Интернет, позволяет автоматизировать процесс обработки релевантной информации с ранжированием смысловой (пертенентной) информации в соответствии с заданной темой, что дает возможность экспертам уйти от ручного последовательного просмотра найденных ресурсов.

Литература:

1. Войскунский В.Г. О построении поисковых характеристик // НТИ. Сер. 2. - 1992. - № 9. - С. 6-9.

2 Барышева О.В., Гиляревский Р.С. О релевантности первичных информационных запросов // НТИ. Сер. 2. - 1995. - № 6. - С. 14-19.