ТРУСОВ Владимир Александрович -
аспирант ГОУ ВПО «Пермский государственный технический университет»
Адрес: 614990, г. Пермь, Комсомольский проспект, 29
e-mail: tva@permcnti.ru

Построение тезаурусов, тематических классификаций и рубрикаторов для поиска информации в распределенных информационных системах

Одним из новых основных понятий, появившихся в результате разработки машинных методов обработки информации, в частности, при переводе с одного языка на другой, поиска научно-технической информации и создания информационной модели предприятия в автоматизированных системах управления, явилось понятие тезауруса информационной системы. Термин «тезаурус» подразумевает совокупность знаний о внешнем мире - это так называемый тезаурус мира Т. Все понятия внешнего мира, выраженные с помощью естественного языка, составляют тезаурус, из которого можно выделить частные тезаурусы путем иерархического деления с учетом соподчинения отдельных понятий или путем выделения частей общего тезауруса мира. Тезаурус в информационно-поисковых системах играет важную роль в поиске нужного документа по ключевым словам. Поэтому построение тезауруса является сложной и ответственнейшей задачей. Но эта задача также может быть автоматизирована.

Классификация в ее наиболее общем определении есть разбиение и упорядочение множеств. Ею называют распределение предметов по классам на основании общего признака, присущего данным явлениям или предметам и отличающего их от предметов и явлений, составляющих другие классы. При необходимости каждый класс может делиться на подклассы. Рубрикатор является особой разновидностью классификации [1]. Поэтому они созданы на основе общих положений:

- научная основа построения классификации;
- отражение современного уровня развития науки;
- наличие системы ссылок и отсылок, а также ссылочно-справочного аппарата (ССА).

Однако рубрикатор является прагматической классификацией, создающейся на основе информационных потоков и потребностей специалистов [2]. В этом его отличие от априорных классификаций, таких как УДК и МПК. Основными функциями классификаций и, в частности, рубрикатора можно назвать следующие:

- тематическое разграничение информационных подсистем;
- формирование информационных массивов по любым признакам;
- систематизирование информационных материалов и изданий;
- текущий и ретроспективный поиск;
- индексирование документов и запросов;
- связь с другими классификационными схемами;
- нормативные функции.

Они строятся путем деления понятий - объектов классификации на основе установленных связей между признаками этих объектов в соответствии с определенными логическими принципами. Признак, по которому производится классификация, получил название основания деления классификации. В классификациях широко используются методы дедукции и индукции для фиксации групп, классов и выявления связей между ними. Это характерно для иерархических классификаций. Глубина классификации (количество уровней иерархии) может быть различной в зависимости от назначения. Одним из широко используемых рубрикаторов является Государственный рубрикатор научно-технической информации (ГРНТИ).

Рубрикатор ГРНТИ разработан так, что возможно его совместное использование с другими классификациями типа УДК и МПК. Универсальная десятичная классификация (УДК) существует более 70 лет, но до сих пор не знает себе равных по широте распространения и используется во многих странах мира. УДК охватывает весь универсум знаний и успешно применяется для систематизации и последующего поиска самых разнообразных источников информации.

Помимо УДК на практике широко используется библиотечно-библиографическая классификация (ББК). ББК построена на принципах логической соподчиненности и представляет классификацию прикладного типа.

Таблица 1

Характеристика рубрикатора ГРНТИ, УДК, ББК и МПК

Наименование	Структура	Принцип расположения делений	Схема построения разделов
УДК	Иерархическая	Отраслевой	От общего к частному
ГРНТИ	Иерархическая	Тематический	Типовая
МПК	Иерархическая	Функционально-отраслевой	От общего к частному
ББК для научных библиотек	Иерархическая	Отраслевой	От общего к частному, по видовому признаку

В Российской Федерации для классифицирования изобретений и систематизации отечественных фондов описаний изобретений используется Международная патентная классификация - достаточно сложная многоаспектная классификация, построенная по функционально-отраслевому принципу. Одни и те же технические понятия могут находиться в МПК или специальных классах (по отраслевой принадлежности) или в функциональных классах (по принципу действия). Отраслевой принцип распределения понятий предполагает классифицирование объектов в зависимости от применения в той или иной исторически сложившейся отрасли техники, технологии.

Сравнительная характеристика рубрикатора ГРНТИ, УДК, ББК и МПК приведена в **таблице 1**.

Таким образом, можно выделить главные отличительные особенности рубрикаторов и классификаторов:

- им свойственен прикладной характер и отраслевая направленность;
- это открытые системы, зависящие от развития науки и техники, потребностей и запросов специалистов;
- неорганичные системы, так как объекты возникают и развиваются в окружающей среде и из нее поступают в них. Элементы способны существовать самостоятельно вне системы. Эта черта тесно связана со второй особенностью;
- минимальным элементом является понятие, связанное со средой. Понятие представляет систему определений;
- между понятиями возникают связи как по «вертикали» (род-вид, целое-часть), так и по «горизонтали» (вид-вид, часть-часть), что свидетельствует об иерархичности систем.

Следовательно, структура и принципы организации классификаций и рубрикаторов делают возможным автоматизировать процесс построения тезаурусов предметной области, используя метод дедукции. Алгоритм построения тезауруса по методу дедукции приведен на **рис. 1**.

Основой для формирования тезауруса является поисковый образ документа, задание или заявка на поиск информации, заполняемая оператором. Следовательно, первым шагом становится исследование и анализ заявки. На первом этапе оператор указывает интересующую тему или проблему, возможные ключевые слова и их синонимы. В результате этого получаем поверхностное представление о предметной области.

Кроме того, формируется тезаурус ключевых слов *KC* по методу дедукции, для чего необходимы:

- массив *KC*, который задает сам пользователь, обозначенный на рис. 1 как *MP*;
- массив *KC*, извлеченный из задания на поиск соответственно *MZ*.

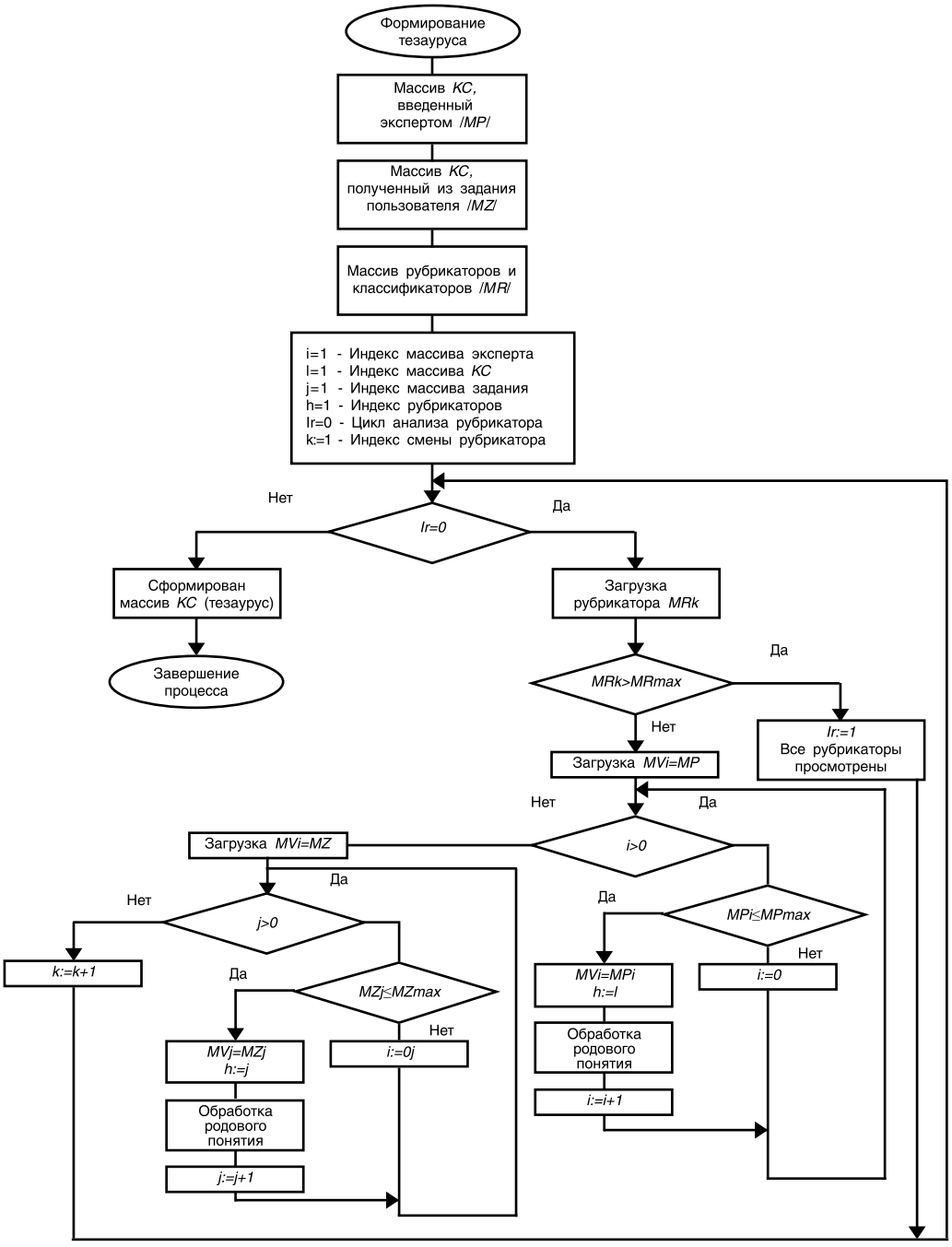

Рис. 1. *Алгоритм построения тезауруса по методу дедукции*

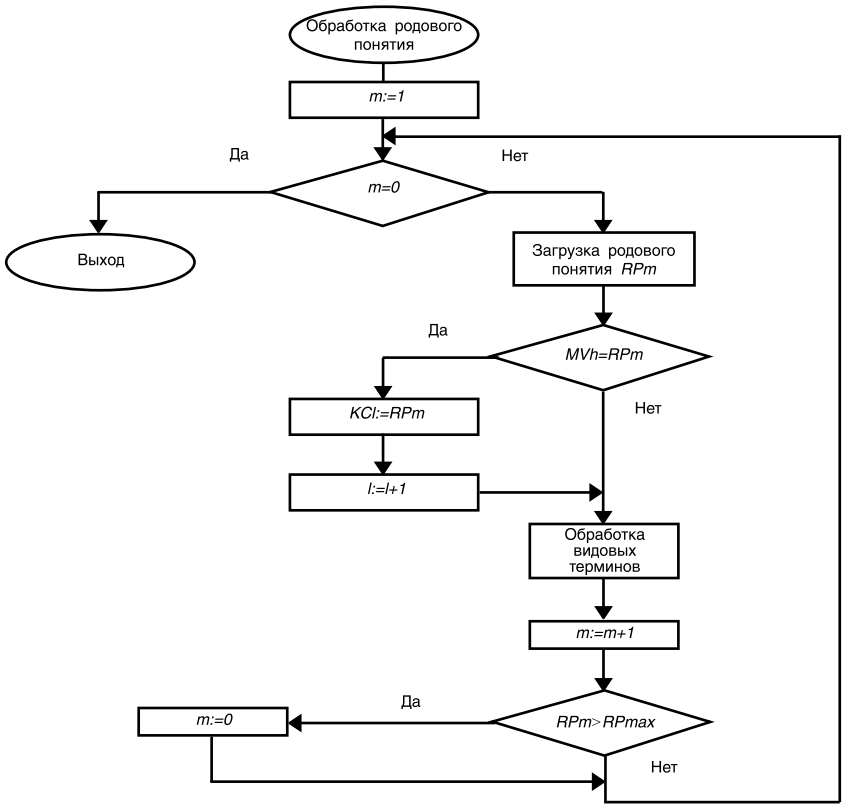

Рис. 2. *Обработка родового понятия*

Однако для более полного и глубинного представления о предметной области используем существующие рубрикаторы и классификационные схемы (ГРНТИ, УДК, ББК, МПК). С целью максимального охвата предметной области необходимо просмотреть все, имеющиеся в наличии. Массив рубрикаторов представляет *MR*. Алгоритм поиска по методу дедукции состоит из двух шагов:

1. Нахождение родовых понятий (**рис. 2**).
2. Нахождение внутри родовых понятий видовых терминов (**рис. 3**).

Загружаем из массива первый рубрикатор и организуем цикл проверки наличия в рубрикаторах *КС*, введенных пользователем. Каждое *КС* ищется в рубрикаторе и сравнивается с родовым понятием или «гнездом», а затем проверяется условие - есть ли ссылка на видовые термины. Если такая ссылка имеется, то *КС* сравнивается с видовыми терминами. В случае если ссылки не обнаружено, переходим к следующему родовому понятию. Когда ключевые слова *КС*, введенные оператором, просмотрены, переходим к массиву *КС*, извлеченных из задания. Процедура проверки аналогична - ищем *КС*, соответствующие родовым понятиям, а затем их ссылки на видовые термины.

Отметим, внутри каждого родового понятия важно просмотреть все имеющиеся видовые термины с целью получения максимального представления о проблемной области. Результатом этих действий является формирование массива ключевых слов *КС*, представляющего собой полный тезаурус, соответствующего заданию на поиск информации или поисковому образу документа.

На базе полного набора поисковых образов документов [3] (обозначим $П=\underset{k}{УП_k}$) можно создать отраслевые тезаурусы и единый классификатор библиотеки. Очевидно, что полный набор *П* сам представляет простейший тезаурус. Однако, используя критерий отбора

$$T_{отр} = \{t_{отр}\} = \underset{t_{отр}}{\cap} П_k , \qquad (1)$$

можем построить отраслевые тезаурусы. При этом множество всех отраслевых тезаурусов образует полный тезаурус

$$T = \{t\} = \underset{k}{UT_k} , \qquad (2)$$

разделы которого могут быть иерархически структурированы в соответствии с требованиями ГОСТов по основным классификаторам (ГРНТИ, УДК, ББК, МПК) или по внутреннему единому классификатору.

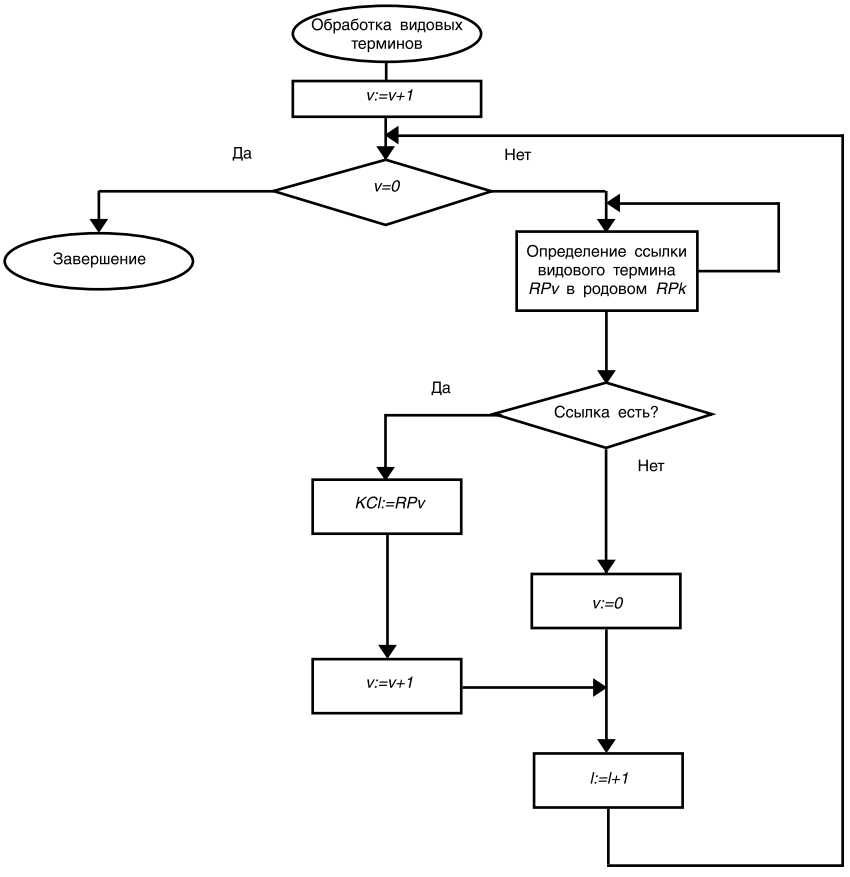

Рис. 3. *Обработка видовых терминов*

Автоматизация процесса построения тезауруса и классификации позволяет максимально облегчить труд оператора, работающего с распределенными информационными ресурсами.

Помимо построения тезауруса, на основе поискового образа документа предложенный подход можно использовать при автоматическом реферировании документа и кластеризации текстов.

Реферирование документов является одной из задач, направленных на обеспечение специалистов-экспертов достоверной информацией, необходимой для принятия управленческого решения о ценности полученных из сети Интернет документов. Реферированием называется процесс преобразования документальной информации, завершающийся составлением реферата, а реферат - это семантически адекватное изложение основного содержания первичного документа, отличающееся экономной знаковой оформленностью, постоянством лингвистических и структурных характеристик и предназначенное для выполнения разнообразных информационно-коммуникативных функций в системе научной коммуникации [4]. Алгоритм реферирования документов представлен на **рис. 4**.

В общем случае алгоритм включает следующие основные этапы:

1. Производим выделение предложений из документа, закачанного из сети Интернет и находящегося в хранилище данных, путем выделения знаков препинания и сохраняем его в массиве.

2. Каждое предложение разбиваем на слова путем выделения разделителей и сохраняем их в массиве, причем для каждого предложения массив разный.

3. Для каждого предложения, для каждого слова этого предложения считаем количество слов в других предложениях (до и после). Сумма повторов для каждого слова (до и после) и будет весом данного предложения.

4. Заданное число предложений с максимальным весовым коэффициентом и выбираем в реферат в порядке появления в тексте.

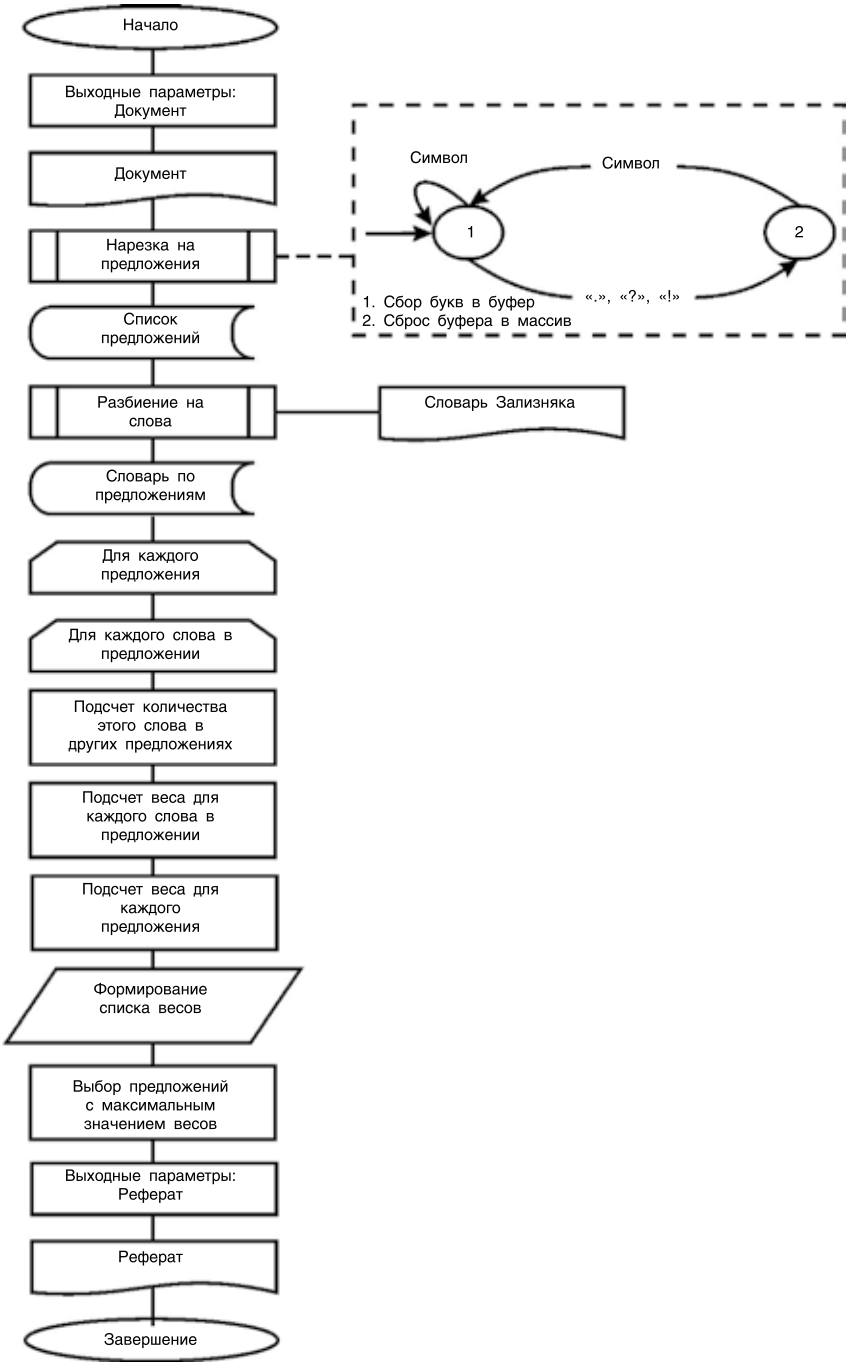

Рис. 4. *Алгоритм реферирования документов*

Предложенная модель построения тезауруса и тематических каталогов информационной системы представляет собой теоретическую основу для автоматизации смыслового поиска и позволяет специалисту-эксперту не только проводить поисковые работы, но и в автоматизированном режиме реферировать документы, полученные в результате поиска в распределенных информационных системах сети Интернет.

Литература:

1. Барушкова Р.И. Классификационные схемы научно-технической информации: Учеб. пособие. - М., 1981. - 80 с.

2. Барушкова Р.И. Рубрикатор как классификационная схема научно-технической информации: Методическое пособие. - М., 1980. - 38 с.

3. Трусов А.В., Бабарыкин Е.П. Оценка границ области тематического информационного запроса в распределенных информационных системах. Материалы Всероссийской (с международным участием) конференции «Информация, инновации, инвестиции», 24-25 ноября 2004 года, г. Пермь / Пермский ЦНТИ. - Пермь, 2004. - С.76-79.

4. Яцко В.А. Логико-лингвистические проблемы анализа и реферирования научного текста. - Абакан: изд-во Хакасского гос. ун-та, 1996. - 128 с.