РОССИЙСКАЯ АКАДЕМИЯ НАУК

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
(ВИНИТИ)

НАУЧНО • ТЕХНИЧЕСКАЯ
ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 3 Москва 2013

ИНФОРМАЦИОННЫЕ СИСТЕМЫ

УДК 001.102-047.44:[004:002.1]

О.Л. Голицына, Н.В. Максимов, О.В. Окропишина, В.И. Строгонов

Онтологический подход к идентификации информации в задачах документального поиска: практическое применение[1]

Предложен подход к определению онтологии и множества операций как инструмента формирования и количественно оцениваемого соотнесения идентифицирующих образов объектов предметной области. Приведены алгоритмы и примеры применения операций над онтологическими представлениями поисковых образов документов и запросов в механизмах поиска в базах научно-технической информации.

Ключевые слова: *онтологии, операции над онтологиями, информационный поиск, идентификация содержания, теория графов*

ВВЕДЕНИЕ

Понятие информации и информационного поиска всегда, так или иначе, связывается с процессом, имеющим неопределенность исхода, и, если это управляемый процесс, – с выбором, который, в свою очередь, использует данные, находящиеся вне ИПС - с наличными знаниями. Неопределенность такого выбора обусловлена последовательными преобразованиями в связываемых посредством ИПС цепочках: «*знания – информация - документ – поисковый образ документа (ПОД)*» и «*проблемная ситуация – задача – запрос - поисковый образ запроса (ПОЗ)*».

Процесс информационного поиска – это построение множества документов, формально соответствующих ПОЗу, посредством процедур, реализующих ту или иную поисковую модель. Здесь необходимо учитывать, что каждое преобразование как в цепочке «знание->ПОД», так и в цепочке «Проблемная ситуация->ПОЗ» представляет собой отображение, причем в пространствах с меньшим разнообразием. Более того, пространства в обеих цепочках для каждого преобразования хотя и подобны (имеют одинаковую природу), но не тождественны. И при этом машинный поиск

[1] Работа выполнена при поддержке РФФИ, грант № 11-09-13128 офи-м-2011-РЖД

как процесс, сводящийся к отбору через сравнение, в общем случае, гипотетического отыскиваемого объекта с объектами, хранящимися в массиве, реализуется не через сравнение самих объектов, а через соотнесение их хорошо структурированных формализованных описаний - поисковых образов[2] [1].

Следует отметить еще одно характерное отличие в формировании и использовании образов в машинной среде ИПС и в сознании человека. Машинные образы создаются обычно в виде статичного набора атрибутов (устойчивой структуры) для отражения *наиболее характерных* свойств. В сознании образы формируются преимущественно вследствие действий и практически не существуют вне связей. Соответственно, машинный отбор образов реализуется по точным критериям, соотносящим исключительно значения (величины) признаков. Поиск же образов в сознании человека производится по ассоциациям (связям), обычно по признаку целевого (предполагаемого) *использования* значения.

В этом смысле поисковые образы, построенные на предложенных онтологических подходах, позволяют работать равно как с признаками, определяющими свойства, так и с признаками, определяющими их взаимосвязь (поведение).

1. ОПЕРАЦИИ НАД ОНТОЛОГИЯМИ В ЗАДАЧАХ ИНФОРМАЦИОННОГО ПОИСКА

Формальное определение онтологии[3] для задач информационного поиска, предложенное в [2], позволяет использовать ее как операционный объект - средство отбора документов с учетом их семантики.

Онтология, построенная по отдельной единице документального потока, может рассматриваться как семантический поисковый образ документа. ПОЗ, в свою очередь, может быть представлен как в традиционной форме (списком дескрипторов), так и в расширенной - с функциональными связями между понятиями (дескрипторами). Соответственно, реализация поисковых механизмов в этом случае основывается на применении операций над онтологиями.

В качестве основных операций над онтологиями в работе [2] были определены операции объединения и пересечения, а также операции проекции и масштабирования.

Традиционные теоретико-графовые операции объединения и пересечения применительно к онтологиям дополнены возможностями сопоставления объектов исходных онтологий с помощью операции тождества не только в функциональной, но и в понятийной и терминологической системах. Операция объединения онтологий, например, может быть использована для построения (и/или последовательного наращивания) онтологии предметной области (ПрО) на базе онтологий отдельных научных исследований, для создания общей онтологии группы исследователей на основе объединения разных точек зрения и т.п. Результат операции пересечения онтологий позволит выявить общее и частное в научных работах, может также служить сигналом заимствования.

Операции проекции и масштабирования (укрупнения, детализации) требуют предварительного построения для исходной онтологии аспектной онтологии или онтологии масштабирования и сводятся к операциям пересечения или объединения построенной и исходной онтологий.

В соответствии с определением операция аспектного представления (рассмотрения, описания) задается функциональной системой $S_f^i = \left\langle M_f^i, A_f^i, R_f^i, Z_f^i \right\rangle$, а результат операции представляет собой пересечение исходной - $O = \left\langle S_f, S_c, S_t, \equiv \right\rangle$ и аспектной - $O_i = \left\langle S_f^i, S_c, S_t, \equiv \right\rangle$ онтологий: $O_{proj} = O \cap O_i$

Операция проекции ориентирована на построение подграфа функциональной системы, отражающего «взгляд» на исходную онтологию с точки зрения некоторого заданного аспекта. Под аспектом рассмотрения (представления) публикации (в данном случае – научной работы) понимается некоторая онтология, фиксирующая объекты рассмотрения и связи (отношения) между ними[4].

Результат пересечения аспектной и исходной онтологий позволит определить контекст заданного аспекта в конкретной работе и далее рассматривать этот контекст с точки зрения реализации поисковых механизмов, алгоритмов автоматической классификации/кластеризации и т.п.

Однако при построении аспектной онтологии следует учитывать, что корректное построение возможно только в случае, когда при задании аспекта каж-

---

[2] Отметим, что при всем разнообразии моделей поиска и мер близости реально (в вычислительной среде) соответствующие алгоритмы сводятся к двоичной логике.

[3] Онтология предметной области авторами формально определена, как $O =< S_f, S_c, S_t, \equiv >$, где

S_f - функциональная система («рабочий интерфейс» онтологии в деятельности субъекта);

S_c - понятийная система (логико-семантический базис онтологии);

S_t - терминологическая система (знаки, используемые для фиксирования онтологии на носителе);

$\equiv$ - операция сопоставления элементов различных систем на уровне знаков, обеспечивающая их тождество в функциональной, понятийной и терминологической системах.

Со структурной точки зрения (для реализации операций над онтологиями) функциональная система может быть представлена помеченным взвешенным направленным мультиграфом, понятийная система - помеченным взвешенным направленным графом, терминологическая система описывается n-связным графом, где каждая компонента связности представляет собой полный граф (эквивалентность), дерево (включение) или результат операции объединения полных графов и деревьев (при наличии общих вершин).

[4] Аспект онтологии может задаваться знаковыми описаниями (1) наборов объектов и/или функциональных отношений. Для последнего случая в [3] типизацию отношений было предложено основывать на функциональной модели деятельности с учетом классических фаз жизненного цикла, которые проходит научная разработка за период своего существования (фундаментальные исследования, прикладные исследования, опытно-конструкторские разработки, серийное производство, применение и утилизация), а также стадий исследования (анализ проблемной ситуации, постановка цели и задачи, исследование аналогов, построение гипотез, моделирование, проверка адекватности, экспериментальное исследование, разработка, тестирование, применение, сопровождение, оценка перспективы развития).

дое функциональное отношение может быть представлено дугой, соединяющей вершины мультиграфа функциональной системы, или функциональные отношения не используются (мультиграф функциональной системы пуст, аспект задается набором знаковых описаний).

Операция масштабирования (укрупнения или детализации) позволяет изменять уровень абстракции представления научного исследования на основе родо-видовых связей понятийной системы онтологии.

Для описания операции масштабирования (укрупнения или детализации) онтологии для исходной онтологии $O=\langle S_f, S_c, S_t, \equiv\rangle$ определяется онтология масштабирования $O_m=\langle S_f^m, S_c, S_t, \equiv\rangle$, а операция масштабирования сводится к построению онтологии $O\cup O_m$. Онтология масштабирования при этом должна содержать знаковые описания объектов, выбранных для масштабирования, и их понятийные деревья. Понятийное дерево объекта представляет собой фрагмент графа понятийной системы с корнем, задаваемым объектом. Видовые связи (связи типа «нижестоящий») заменяются функциональным отношением «является частью/частным случаем».

После объединения исходной онтологии и онтологии масштабирования связи «является частью/част-ным случаем» сворачиваются к конечному (в случае укрупнения) или к начальному (в случае детализации) объекту.

Операция масштабирования может применяться к исходной онтологии и в отсутствии онтологии масштабирования (например, если объектам масштабирования не нашлось тождественных в понятийной системе). В этом случае операция укрупнения/детализации сводится к сворачиванию отношений «является частью/частным случаем» в исходной онтологии.

Масштабирование онтологий может применяться для приведения их к одному понятийному контексту перед использованием операций объединения, пересечения и проекции.

2. ПРИМЕРЫ ИСПОЛЬЗОВАНИЯ ОПЕРАЦИЙ ПРОЕКЦИИ И МАСШТАБИРОВАНИЯ

Рассмотрим операцию аспектного представления, где для построения аспекта используется ПОЗ. Возможны следующие случаи.

1. ПОЗ задан в традиционной форме в виде списка дескрипторов, что соответствует ситуации, когда аспект задается набором знаковых описаний.

Для мультиграфа аспектной онтологии $MG_f^i=\langle V_f^i, \varnothing\rangle$ при этом необходимо сформировать множество вершин, в которое должны войти дескрипторы ПОЗ.

Для вычисления соответствия документа поисковому запросу выполняется операция пересечения семантического ПОД и аспектной онтологии. Множество вершин результирующей онтологии формируется по правилам операции пересечения (с использованием понятийной и терминологической систем), а множество дуг - из дуг, связывающих вершины. Результат формального применения операции пересечения может быть при необходимости изменен в соответствии со следующим правилом: если результирующий мультиграф получился несвязным, то в него может быть включено минимальное подмножество дуг и вершин, дополняющих его до связного мультиграфа.

Рассмотрим на примере применение операции аспектного представления при поиске (отборе документа по запросу). На рис. 1 представлен фрагмент мультиграфа функциональной системы онтологии, построенной по тексту диссертации [4] средствами информационно-аналитической системы xIRBIS [5].

Формулировка поискового запроса включает термины «теория Абрикосова» и «магнитные вихри», т.е. множество вершин мультиграфа аспектной онтологии V_f^i={«теория Абрикосова», «магнитные вихри»}. При этом в качестве общей понятийной системы используем тезаурус INIS.

Результат выполнения операции проекции семантического ПОД на аспектную онтологию представлен на рис. 2. Вершины «сверхпроводник» и «сверхпроводимость» попали в результат как ассоциативные в тезаурусе дескриптору «теория Абрикосова» (при пересечении с учетом понятийной системы), а вершина «пиннинг магнитных вихрей» - в процессе использования терминологической системы как словосочетание, включающее дескриптор «магнитные вихри». Вершины «смешанное состояние», «дефекты», «высокотемпературный слоистый сверхпроводник» вместе с дугами включены для соединения вершин из множества пересечения (как минимальное подмножество дуг и вершин, дополняющих мультиграф до связного).

Построенное аспектное представление документа далее может быть использовано как при вычислении критерия смыслового соответствия документа запросу, так и в процедуре кластеризации выдачи.

2. ПОЗ помимо дескрипторов содержит функциональные связи. В этом случае аспектная онтология может быть построена только если функциональные связи заданы между дескрипторами запроса (т.е. может быть сформирован мультиграф $MG_f^i=\langle V_f^i, X_f^i\rangle$ с непустыми множествами вершин и дуг). Тогда аспектное представление документа строится по правилу операции проекции. При необходимости аспектная онтология может быть масштабирована.

Например, для поискового запроса «Дефекты как свойство сверхпроводника» может быть построена аспектная онтология, представленная на рис. 3.

Операция проекции дает в результате пустой граф, поэтому над аспектной онтологией выполняется операция масштабирования (детализации). В качестве онтологии масштабирования рассмотрим фрагменты тезауруса – видовые деревья с корнями в вершинах – «дефекты» и «сверхпроводник». Видовые (нижестоящие) отношения понятийного уровня заменяются дугами «является частью/частным случаем». Полученная для дескриптора «сверхпроводник» (так как дескриптор «дефекты» не имеет нижестоящих) онтология масштабирования представлена на рис. 4.

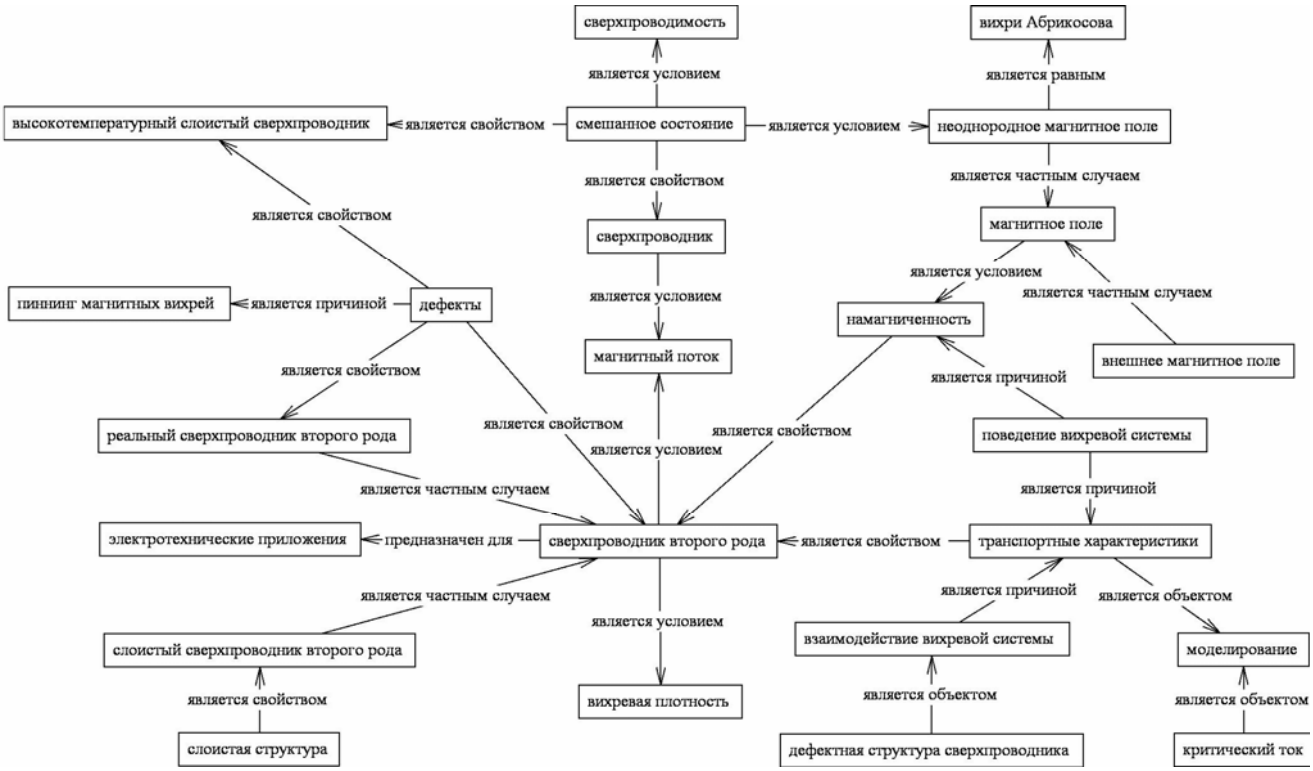

Рис. 1. Фрагмент мультиграфа функциональной системы онтологии

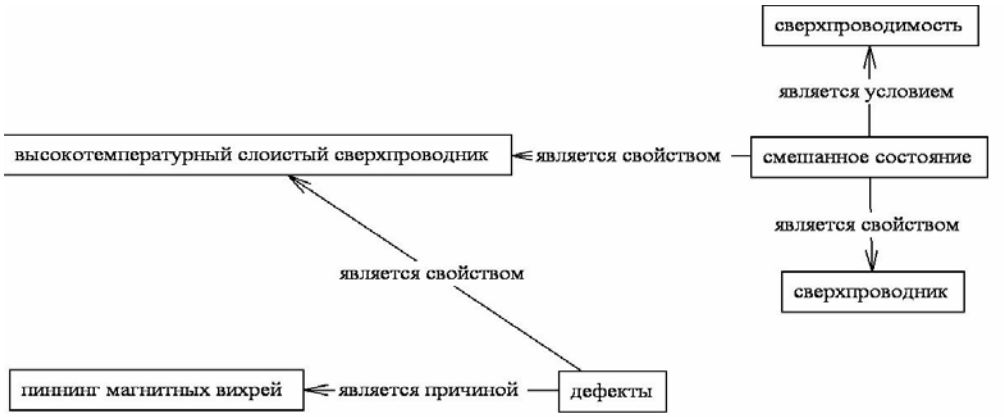

Рис. 2. Результат проекции по запросу «Теория Абрикосова и магнитные вихри»

Рис. 3. Мультиграф функциональной системы аспектной онтологии для запроса «Дефекты как свойство сверхпроводника»

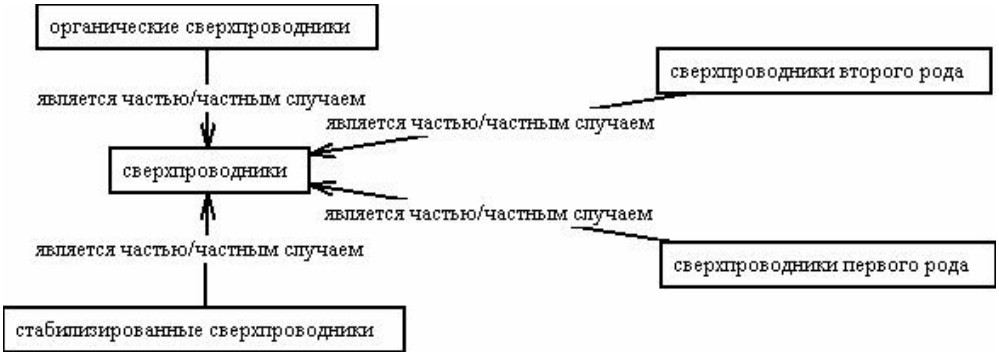

Рис. 4. Онтология масштабирования

Результат объединения аспектной онтологии и онтологии масштабирования представлен на рис. 5.

Следующий шаг детализации – стягивание вершин, инцидентных дугам «является частью/частным случаем», к вершине – началу дуги (рис. 6). Применение в операции проекции детализированной онтологии дает результат, представленный на рис. 7.

Однако по поисковому запросу не всегда может быть построен мультиграф, т.е. функциональные связи представляют собой дуги, для которых не заданы конечные вершины. В этом случае для получения результата соотнесения ПОЗ и ПОД может быть использована операция масштабирования с онтологией масштабирования, построенной на дескрипторах запроса.

Рассмотрим поисковый запрос «Свойства сверхпроводника». Дескриптор «сверхпроводник» задает объект запроса, а дескриптор «свойство» соответствует функциональному отношению «является свойством».

При масштабировании (в данном случае – укрупнении) исходной онтологии с использованием онтологии масштабирования, представленной рис. 4, на первом шаге (объединение онтологий) вершина «сверхпроводник второго рода» будет соединена с вершиной «сверхпроводник» дугой «является частью/частным случаем». На следующем шаге укрупнения происходит стягивание вершин, инцидентных дугам «является частью/частным случаем», к вершине – концу дуги. При этом в вершину «сверхпроводник» будут последовательно стянуты вершины «реальный сверхпроводник второго рода», «слоистый сверхпроводник второго рода» и «сверхпроводник второго рода» (рис. 8)

На следующем шаге строится подграф, включающий вершину «сверхпроводник» и все маршруты от этой вершины, состоящие из дуг «является свойством». Полученная таким образом онтология - результат проекции по запросу «Свойства сверхпроводника», представлена на рис. 9. .

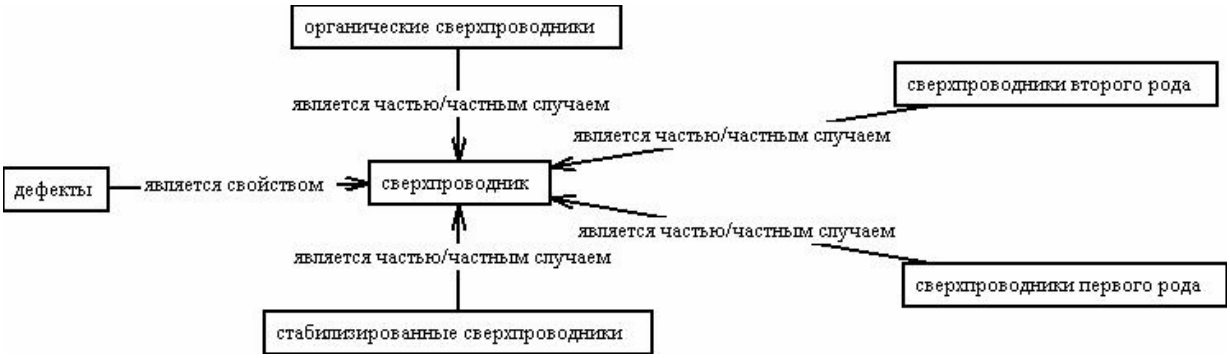

Рис. 5. Результат объединения аспектной онтологии и онтологии масштабирования

Рис. 6. Детализированная онтология масштабирования

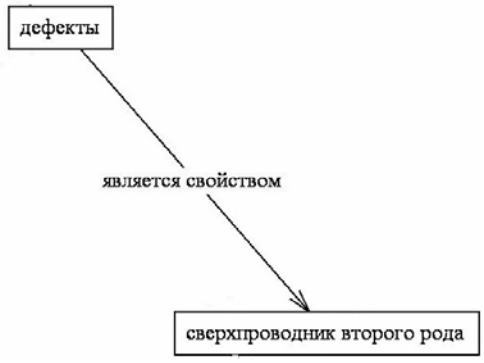

Рис.7. Результат проекции по запросу «Дефекты как свойство сверхпроводника»

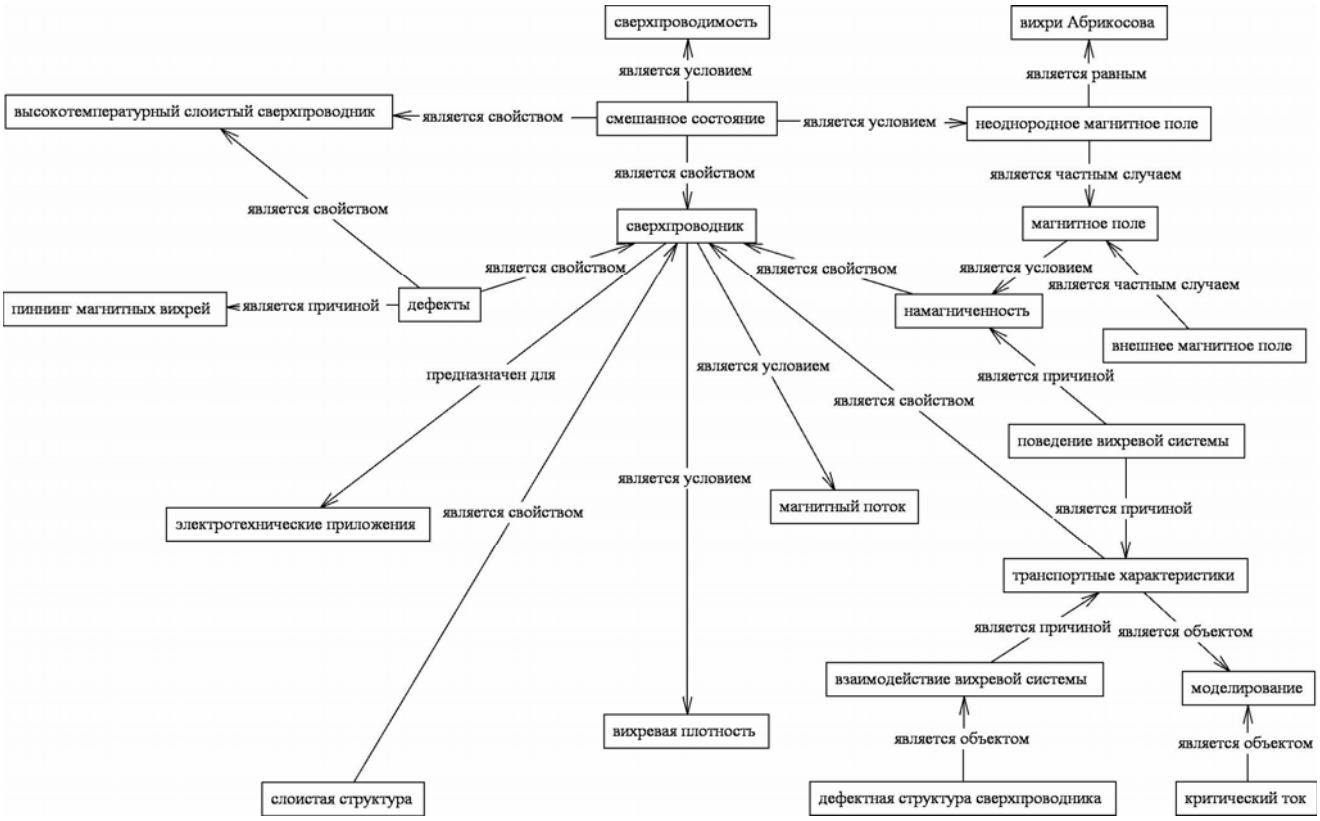

Рис. 8. Укрупнение онтологии

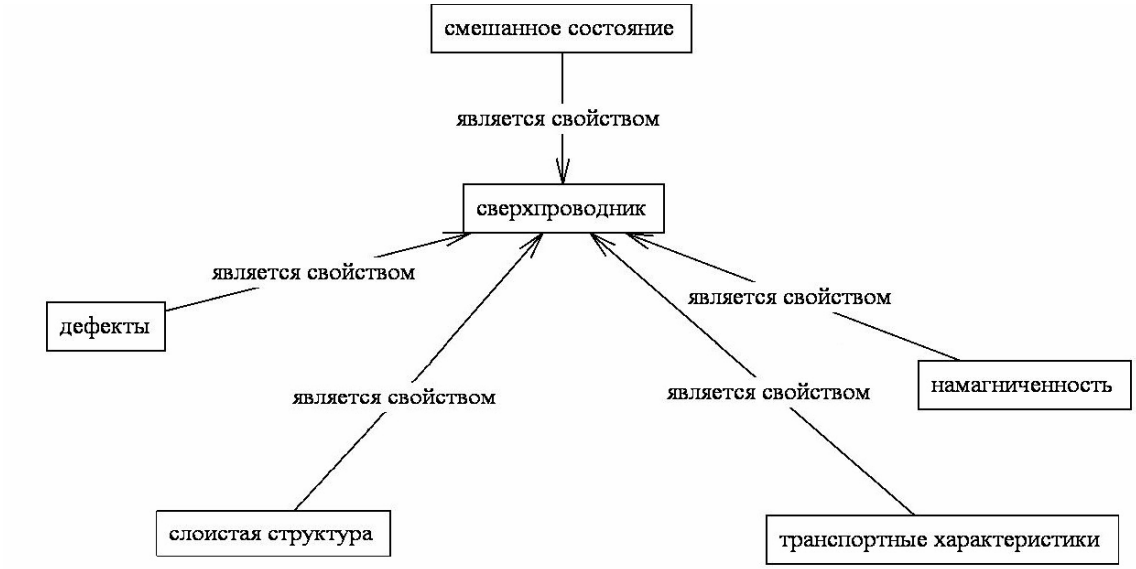

Рис. 9. Результат проекции по запросу «Свойства сверхпроводника»

3. ПРАКТИКА ИНФОРМАЦИОННОГО ПОИСКА

«Нечеткость» отображений, отмеченная во Введении, означает, что неточность любого из соответствий должна быть компенсирована за счет избыточности выдачи, что, соответственно, приведет к увеличению полноты выдачи и обычно к уменьшению точности. Это может быть достигнуто следующим образом: 1) реформулирование и обогащение выражения запроса, в частности с использованием других терминологических систем; 2) использование нескольких механизмов поиска, в том числе с разной жесткостью критерия и основанных на технологии обратной связи по релевантности [1].

Первый путь – лингвистический – нахождение множества имен понятий и, соответственно, терминов ИПЯ, образующих в некотором заданном аспекте класс условной эквивалентности для каждого из исходных объектов запроса. Другими словами - это

формирование возможных вариантов терминологических конструкций, выражающих существо информационной потребности (ИП) в этом аспекте. Тезаурусы и семантические словари, используемые для такого расширения, в известной степени обеспечивают конкретизацию семантики терминов (но в рамках этой конкретной области знаний), однако средства такого рода не могут отражать семантику реальной ИП: в общем случае, вследствие динамизма процесса познания, трудно предусмотреть будущие возможные смыслы терминов при введении их в понятийную систему.

Процедура реформулирования ПОЗ по обратной связи по релевантности позволяет приводить пользовательское представление в соответствие с терминологической системой ИПС. Она реализует лингвистические особенности *использования* языка, представляемые статистическими характеристиками, построенными на основе частотных показателей термина в сопоставлении с его семантикой. Значение термина (как лингвистической переменной) в рамках более крупных конструкций, таких как предложение или документ, определяется пользователем достаточно точно (хотя и субъективно) и обычно без явного использования метаинформации. Менее точно и полно может быть определен смысл термина в наборе документов. Но именно использование в поисковых механизмах аспектной проекции поискового образа, содержащего меньшее количество (чем в исходном полном ПОДе), но более точных терминов (за счет большего числа слов, входящих в термин), позволяет существенно увеличить точность выдачи.

Это объясняется тем, что с точки зрения, представляющей информацию как данные во взаимосвязи с контекстом [1], семантические связи, порождаемые уже совокупностью лингвистических отношений такого (составного) термина, формируют практически однозначный контекст.

Существенным фактором здесь является и то, что практика пользовательских интерфейсов представления поисковых запросов преимущественно в вербальной форме ориентирована, по существу, на фактографический вид поиска: в поисковом условии в линейном лаконичном виде приводятся в основном артефакты (свойства, величины, имена, термины и т.п.), а не связи, которые собственно и формируют контекст и, в итоге, конкретный смысл. Другое известное интерфейсное (правильнее, технологическое) решение - кластерные технологии поиска (в частности, реализации технологии реформулирования запроса по обратной связи по релевантности) дают по сравнению с вербальной процедурно эффективные решения. Ввод с клавиатуры хорошо контекстно определенных, но длинных словосочетаний, практически невозможен, в то время как в кластерных технологиях система формирует возможные контексты в виде кластеров терминов, выбираемых пользователем из списка статистически значимых терминов для релевантных документов.

В этом смысле поисковые образы, построенные на предложенных онтологических подходах, расширяют возможности обоих технологий, позволяя работать равно как с признаками, определяющими свойства, так и с признаками, определяющими их взаимосвязь, а операции над такими описаниями обеспечивают их преобразование и взаимное сопоставление через приведение к целевому контексту, определяемому целями субъекта поиска и конкретной предметной области.

ЗАКЛЮЧЕНИЕ

Использование предложенного определения онтологии и множества операций как инструмента формирования и количественно оцениваемого соотнесения образов объектов предметной области позволяет перейти от вычислительной (корректной) задачи к задаче построения самосогласованной системы - формированию диалектически взаимосвязанных пространств абстрактных и конкретных объектов ПрО, коррелированно отражающих в пространство знаков содержание знания в двух формах: знания состоявшегося (документов) и гипотетического (запросов).

Такой подход отражает существо диалектики когнитивного процесса: проблемная ситуация, обнаруживаемая во взаимодействиях объектов реальности, приводит к построению модели – абстракции, оперирующей идеализированными объектами, последующие построения на множестве которой позволяют синтезировать новые объекты и взаимодействия реальности, тем самым изменяя её.

В задачах информационного поиска такой подход позволит в автоматизированных процессах динамического реформулирования и соотнесения поисковых образов запросов и документов не только учитывать разнообразие точек зрения, но и выявлять потенциальные направления развития предметной области.

СПИСОК ЛИТЕРАТУРЫ

1. Голицына О.Л., Максимов Н.В. Модели информационного поиска в контексте поисковых задач // НТИ. Сер.2. - 2011. - № 2. - С.1-12.
2. Голицына О.Л., Максимов Н.В., Окропишина О.В., Строгонов В.И. Онтологический подход к идентификации информации в задачах документального поиска // НТИ. Сер. 2. – 2012. - № 5. – С. 1-9.
3. Максимов Н.В., Окропишин А.Е., Окропишина О.В., Передеряев И.И. Использование технологии автоматизированного формирования понятийной структуры предметной области научного исследования в задачах управления научными кадрами // Вестник РГГУ. Сер. «Управление». - 2011. - № 4 (66) – С. 175-185.
4. Одинцов Д. С. Моделирование транспортных характеристик высокотемпературных сверхпроводников: дис. … канд. физ.-мат. наук. - М.: МИФИ, 2008.
5. Максимов Н.В., Голицына О.Л., Окропишин А.Е., Окропишина О.В. Подсистема аналитической обработки документальной информации. Свидетельство о государственной регистрации программы для ЭВМ №2011611694 от 22.02.2011г.

Материал поступил в редакцию 26.12.12.

Сведения об авторах

ГОЛИЦЫНА Ольга Леонидовна – доцент, кандидат технических наук, доцент каф. Системного анализа Национального исследовательского ядерного университета «МИФИ»

E-mail: OLGolitsina@YANDEX.RU

МАКСИМОВ Николай Вениаминович – доктор технических наук, профессор каф. Системного анализа Национального исследовательского ядерного университета «МИФИ»

E-mail: NV-MAKS@YANDEX.RU

ОКРОПИШИНА Ольга Владимировна – инженер каф. Системного анализа Национального исследовательского ядерного университета «МИФИ»

E-mail: doomguard@YANDEX.RU

СТРОГОНОВ Владимир Иванович - доктор технических наук, Зам. руководителя Центра перспективных фундаментальных и прикладных исследований ОАО «Научно-исследовательский и проектно-конструкторский институт информатизации, автоматизации и связи на железнодорожном транспорте»

E-mail: Strogonov_vi@mail.ru