

Онтологический подход к идентификации информации в задачах документального поиска

Предложен подход к определению онтологии и множества операций как инструмента формирования и количественно оцениваемого соотношения идентифицирующих образов объектов предметной области в условиях диалектической взаимосвязи предметного, понятийного и знакового пространств. Онтологическое представление образа объекта в вычислительной среде соответствует объектно-ориентированному подходу и включает не только свойства, но и поведение. На практике такой подход позволит автоматизировать процессы динамического реформулирования и соотношения поисковых образов запросов и документов на основе их приведения к общему понятийно-терминологическому контексту.

Эффективность поиска и адекватность передачи смысла в системе кругооборота информации, обеспечивающей расширенное воспроизводство знаний, связаны с выбором или введением специальной терминологии. Одним из перспективных направлений в области формализации знаний, которое дает возможность их компьютерной обработки, являются онтологии.

Наиболее общее определение дано в [Gruber]: «Онтология – это спецификация концептуализации», и существенно уточнено в [Van Heijst], где онтология определяется как «спецификация концептуализации на уровне эксплицитных знаний, зависящая от предметной области или задачи, для которой она предназначена». Под «концептуализацией» понимается строгое описание системы понятий, объектов и других сущностей, а также отношений, связывающих их друг с другом. С другой стороны, концептуализация – это абстрактное, и т.о. упрощенное видение мира, который мы хотим представить для достижения какой-то цели. Таким образом, концептуализация расчленяет целостную область знаний, выделяет из этой области отдельные объекты и отношения, то, что с точки зрения конкретной цели будет адекватным операбельным объектом, заменяющим саму предметную область (ПрО) в процессе исследований.

В общем виде структура онтологии представляет собой набор элементов четырех категорий: понятия, отношения, аксиомы, отдельные экземпляры. Понятия рассматриваются как концептуализации класса всех представителей некой сущности. Понятия могут быть связаны различного рода отношениями. Аксиомы задают условия соотношения категорий и отношений.

Однако в «инженерной» области онтологии сильно различаются. Например, с точки зрения предмета концептуализации выделяют прикладные онтологии, онтологии области знания, общие (родовые) и репрезентационные онтологии (онтологии метауровня) [Guarino]. Отдельно выделяют лингвистические онтологии (проекты WordNet, MikroKosmos, РуТез и другие), фиксирующие понятия (слова) вместе с их языковыми свойствами и отношениями (синонимия, гипонимия и т.п.). Существует и два подхода к концептуализации: интенциональный и экстенциональный. Экстенциональный [Genesereth] подразумевает, что каждое понятие и отношение может исчерпывающе описываться перечислением индивидуальных сущностей. При интенциональном подходе предлагается [Guarino] идентифицировать понятия не через их перечисление, а через их внутренние свойства и характеристики.

Показателен подход, предложенный в [David De Roure], где онтология – это файл с метаданными, предназначенными для аннотирования содержания ресурса и сервисов в нескольких аспектах, который должен включать следующие типы метаданных:

- 1) описание объектов, их свойств и отношений между ними (Domain ontologies);

- 2) описание задач и процессов, их свойств и отношений (Task ontologies);
- 3) описание атрибутов знания (Quality ontologies);
- 4) характеристику значимости контента (Value ontologies);
- 5) историю обращений отдельного субъекта к информационным ресурсам (Personalization ontologies);
- 6) описание причин накопления, оценок использования (Argumentation ontologies).

Определение, конструктивное с т.з. формализации, предложено в [Бениаминов]: «Онтология – это набор определений (на формальном языке) фрагмента декларативных знаний, ориентированный на совместное многократное использование различными пользователями. В онтологии вводятся термины, типы и соотношения (аксиомы), описывающие фрагмент знания».

В задачах разметки текстов учет семантических категорий, описанных в онтологии, позволяет сделать идентификацию более точной, уменьшить неоднозначность. Такая семантическая разметка в дальнейшем позволяет проводить семантический анализ текста, различные статистические исследования, извлекать межъязыковые соответствия [Мудрая].

В целом основная цель создания онтологий - сделать возможным *разделение (sharing)* и *повторное использование (reuse)* знаний и, в том числе, обеспечить *интероперабельность* между несовместимыми инструментальными средствами в среде моделирования деятельности человека. Практическое применение онтологий - это соглашение об использовании общего словаря в согласованной и логичной манере [M. Uschold].

В данной статье с общесистемных позиций будет рассмотрен семиотический подход к определению онтологии, позволяющий учесть ту важную для идентификации обрабатываемых объектов особенность машинной обработки, что как абстрактные, так и конкретные объекты, как классы, так и их экземпляры в вычислительной среде будут представлены знаками.

Введение в задачу идентификации ПрО

Любой целенаправленный процесс синтеза нового знания как любых новых объектов явно или неявно включает отдельный подпроцесс - выбор из множества доступных только нужных «строительных блоков». Причем выбор здесь – это не столько собственно извлечение конкретного объекта из среды, сколько способ и процесс отождествления объекта либо с образцом (в частности, с аналогом), либо с «нишей», им заполняемой. Однако в большинстве процессов разумной деятельности выбор реализуется не по приведенной «натурной» схеме, а опосредовано – через информационное взаимодействие, где сопоставляются образы: образ объекта – кандидата на отбор с образом целевого объекта (возможно, существующего только гипотетически). Такой подход позволяет реализовать эффективную (по используемым ресурсам) процедуру: объект с потенциально бесконечным разнообразием свойств и связей заменяется в операции сопоставления компактным хорошо формализованным идентификатором, свойства которого представлены в форме, адекватной операционной среде.

Такого рода образами обычно являются описания объектов или процессов (научные, технические публикации, математические модели, проектная документация и т.п.), в той или иной форме представляющие те или иные свойства и поведение объекта с полнотой и точностью, обусловленной характером решаемой практической задачи.

При этом и в человеческом сознании, и в вычислительной среде чаще сравниваются не непосредственно описания, а их «производные» по тем аспектам рассмотрения

предметной области, которые отвечают конкретной задаче потребителя – субъекта, формирующего новое знание¹. Такое положение обусловлено нижеследующим.

1) С точки зрения Общей теории систем [Урманцев] знания, как и описания любых объектов, по своей природе системны: состав элементов, их свойства и взаимосвязи для каждого случая (аспекта рассмотрения, разных задач и т.д.) будут определяться своим законом композиции. То есть, для отдельной ПрО может быть определено множество систем $S^i = \langle M^i, A^i, R^i, Z^i \rangle$, где M^i - множество объектов ПрО, A^i - множество характеристических атрибутов, R^i - множество отношений, Z^i - закон композиции, $i = 1..n$.

2) Любое описание (как информационный объект² - материальная форма передачи знаний) является не более чем некоторой моделью рассматриваемой ПрО, отражающей только те её свойства и поведение, которые являются существенными для конкретной задачи и обстоятельств, связанных с созданием, целями и методами использования.

3) Описания имеют семиотическую природу, т.е. представлены знаками, конкретные значения которых определяются некоторой конкретной концептуальной схемой ПрО. В целом построение описания (операционного образа) отвечает схеме, представленной на рис. 1.

4) Описание, чтобы быть адекватно воспринятым и понятым, должно быть представлено понятиями той степени общности/детальности и зафиксировано теми терминами, которые будут иметь одинаковые значения как для генератора, так и для приемника сообщения.

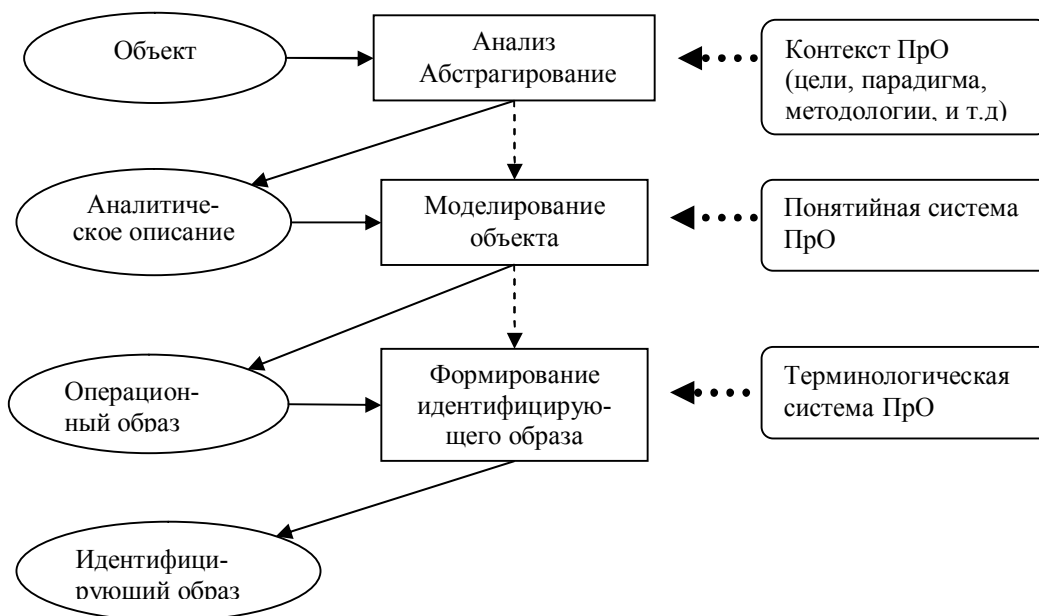


Рис. 1. Построение идентифицирующего образа

¹ Необходимо четко различать описание (зафиксированное на носителе) собственно знания и его внешнее представление (описательный, поисковый образ), как по назначению, так, соответственно, построению и использованию. Если описание представляет существо объекта «изнутри» (как устроен, как достигаются целевые показатели и т.п., что обеспечивает возможность изменения, иного применения), то идентификатор (поисковый образ) представляет объект «извне» - основными свойствами, целевыми показателями (наименованиями и значениями), что обеспечивает его узнавание путем соотнесения с соответствующими характеристиками и, таким образом, выделение из множества других.

² Знание определяется как редукция информации, представляемой как суперпозиция возможных состояний информационного объекта в различных предметных областях. Информация выполняет коммуникационные функции, обеспечивающие взаимодействия в предметной области, и собственные, реализующие её действительность. С точки зрения механизма взаимодействия информация, как структурированный не атомарный объект, должна иметь «интерфейсные» элементы, сопряженные с объектами взаимодействия [Максимов].

Отметим, что успешность восприятия может быть оценена состоянием, когда термины полученного сообщения имеют эквивалентное отображение в понятийно-терминологической системе приемника и образуют логически непротиворечивую структуру; успешность понимания может быть оценена состоянием, когда с помощью построенной по полученному сообщению понятийной системы можно воспроизвести гипотетический объект, обладающий предполагаемыми свойствами и поведением (в частности, свойствами объекта - оригинала).

Процесс восприятия-понимания по существу является обратным по отношению к представленному на рис. 1: производится восстановление объекта по его образу. Очевидно, что в простейшем случае такое восстановление построено по принципу замещения знаков и/или их комбинаций соответствующими «конструкционными» блоками, выбираемыми из хранилища (в частности, памяти субъекта) в соответствии с контекстом, образуемым системой понятий и знаков, а также принятой парадигмой, категориальной системой, конкретными целями и задачами.

Достаточно простая и очевидная схема, приведенная на рис. 1, позволяет, тем не менее, явно выделить принципиальное отличие организации когнитивных процессов в живых системах от искусственных систем сохранения и управления знаниями. Живым системам свойственны целеполагание и целеустремленность. При этом цель, формируемая в среде субъекта после получения внешнего воздействия (сигнала), существенным образом будет определяться контекстом³, образуемым тесно связанными со средой знаниями субъекта: сигнал не будет воспринят, если у субъекта не возникнет ассоциаций с его личным знанием. То есть, данные, представляющие внешнее воздействие, изначально будут связаны с определенным контекстом (хотя, скорее всего, разного даже для разных когнитивных состояний субъекта) и будут обрабатываться совместно, включая, если необходимо, и изменение контекста.

Необходимо отметить особенности, сказывающиеся на условиях формирования контекста в системах сохранения знаний в разные исторические периоды. Для периода индустриального этапа развития общества (собственно и породившего информационные системы и информационную деятельность как класс) было характерно сравнительно небольшое число глобальных проектов и приоритетных отраслей, каждое из которых могло централизованно управляться и реализовываться в рамках одной (зачастую специально создаваемой) организационной структуры, включающей и целевую подготовку профессиональных кадров. Это обеспечивало контролируемое «позадачное» формирование контекста и в сознании отдельного субъекта – будущего решателя отраслевых задач, и в общественном сознании - общего контекста, который, явно или неявно, определял развитие направления. Современное существенное увеличение объемов опубликованного знания, диверсификация исследований, преобладающая децентрализация управления привели к необходимости управления собственно контекстом: уже недостаточно перечислить термины темы исследования – необходимо уточнить их значение; недостаточно назвать зада-

³ Используя терминологию автоматизированных ИПС можно сказать, что контекст является запросом на отбор данных из внешней среды. То есть «информационный ресурс» субъекта формируется «по-задачно»: получаемые данные непосредственно связываются с их целевым использованием. Ресурсы же хранилищ обобществленного знания (библиотеки, АИПС), напротив, формируются по принципу накопления документов, и, в основном, с целью их последующего расширенного использования для построения нового знания, возможно, в областях далеких от изначальной. То есть, можно сказать, что создание нового знания всегда локализовано и конкретно обусловлено, в то время, как его использование всегда вариантно и рассредоточено. При этом для обеспечения широты и многоаспектности использования, в том числе для решения еще не поставленных задач, знание, целостное исторически и методологически, публикуется «по частям» (в соответствии с этапами работ, ограничениями, свойственными видам публикаций, правовыми аспектами и т.д.) и приводится в форме с максимальным абстрагированием (что, в принципе, должно увеличивать его универсальность). И, более того, для обеспечения полноты поиска для решения будущих задач справочно-поисковые средства хранилищ (картотеки библиотек, поисковые образы документов в АИПС) представляют эти знания на достаточно общем уровне и в унифицированной форме.

чу – необходимо привести цели, условия и парадигму её решения. Исходя из приведенных особенностей развития коммуникационных процессов в когнитивных системах, можно сказать, что основой для обеспечения эффективности передачи и расширенного воспроизводства знаний является согласованность и устойчивость используемых понятийно-терминологических систем.

В этом смысле онтологический подход к формированию образа (описания) знаний представляется конструктивным, поскольку позволяет идентифицировать не только сам объект (информацию, знания, предметы и т.п.) на основе его свойств, но и его «поведение» в концептуальном и историческом пространствах.

Такой подход технологически не является принципиально новым: еще в первых поколениях автоматизированных ИПС для построения поискового образа - идентифицирующего описания объекта (документа, описывающего предмет, процесс и т.д.) использовались указатели роли и связи, что позволяло в дескриптивной форме отражать существо объекта, в том числе, с разных точек зрения. Сформированный образ вполне адекватно отражал и свойства, и связи ПрО, однако использованный контекст описания оставался вне образа. То есть, в этом случае не учитываются специфические *особенности и динамика взаимообусловленности предметного, понятийного и знакового пространств*. Современные же условия, как отмечалось выше, предопределяют, что отдельная ПрО может быть представлена разными понятийными и терминологическими множествами (свойство системности) и с разной степенью полноты и детальности (свойство голографичности информации [Бониц]). Отметим, что и операции над такими описаниями должны обеспечить их преобразование и взаимное сопоставление через приведение к целевому контексту, определяемому выбираемыми каждым отдельным субъектом целями и характером решения, и соответствующими обстоятельствами конкретной ПрО. Соответственно, представление образа в вычислительной среде соответствует объектно-ориентированному подходу и будет включать не только данные (свойства), но и методы (поведение) – способы построения, использования и т.д.

Отметим, что термин «онтология» в дальнейшем изложении будет использован в узком смысле поисковых задач (а не задач искусственного интеллекта) – как адаптивный композиционный идентификатор, отражающий основные свойства объекта в сопоставимой форме, и не предполагающий построение объектов, обладающих новизной.

Определение онтологии ПрО

Рассматривая онтологию ПрО с выше определенной узкой точки зрения (идентификации содержания документа при информационном поиске) будем представлять её как совокупность трех систем – функциональной, понятийной и терминологической.

Функциональная система («рабочий интерфейс» онтологии в деятельности субъекта) представляет объекты ПрО и отношения между ними средствами знакового уровня. Эти отношения имеют функциональную окраску, т.к. определяют существенные с точки зрения задач пользователя [Максимов, 2011] способы и характер совместного существования и использования объектов.

Однако функциональная система не может рассматриваться вне связи со сложившейся в ПрО на *текущий момент времени* системой понятий, которая обычно фиксируется в форме тезаурусов, рубрикаторов, классификационных схем (логико-семантический базис онтологии). В понятийной системе объектами (представленными также средствами знакового уровня) являются устойчивые понятия ПрО, а набор отношений ограничен родо-видовыми и ассоциативными.

Терминологическая система в онтологии отражает свойства ЕЯ на уровне знаков - терминов ПрО, которые могут быть связаны отношениями эквивалентности (синонимии) и включения (образования словосочетаний). В качестве термина выступает отдельное слово или словосочетание ЕЯ (а также шифр классификации), которое может быть использовано для описания понятия или объекта в рамках заданной ПрО.

Для определения (установления) взаимосвязи этих систем (а также и их влияния друг на друга в процессе развития) может быть положена простая операция тождества элементов на знаковом уровне.

Таким образом, некоторая онтология ПрО (с точки зрения задач информационного поиска) формально может быть определена, как:

$$O = \langle S_f, S_c, S_t, \equiv \rangle, \text{ где}$$

S_f - функциональная система;

S_c - понятийная система;

S_t - терминологическая система;

\equiv - операция сопоставления⁴ элементов различных систем на уровне знаков, обеспечивающая их тождество в функциональной, понятийной и терминологической системах.

Определенная таким образом онтология также обладает свойствами системы (с т.з. ОТС), т.е. для отдельной предметной области (в том числе и отдельного её элемента) может быть построено столько онтологий, сколько может быть аспектов рассмотрения ПрО⁵, т.е. $O^i = \langle S_f^i, S_c^i, S_t^i, \equiv \rangle$. Однако с целью упрощения выкладок, если не будет явной необходимости, мы будем опускать соответствующий индекс.

Отметим, что в качестве функционального (и терминологического) компонента может быть и система понятий ПрО. В этом случае её понятийной системой будет система метапонятий (это отвечает, например, определению «метаонтологии» [Nirenburg], где на верхнем уровне есть лишь три категории: «объект», «процесс» и «роль»).

Рассмотрим компоненты, входящие в приведенное выше определение онтологии.

Функциональная система определяется как:

$$S_f = \langle M_f, A_f, R_f, Z_f \rangle, \text{ где}$$

M_f - множество знаковых описаний объектов ПрО;

A_f - множество характеристических атрибутов;

R_f - множество функциональных отношений;

Z_f - закон композиции, в соответствии с которым выбрано конкретное системное основание $\{M_f, A_f, R_f\}$.

Со структурной точки зрения⁶ функциональная система может быть представлена помеченным взвешенным направленным мультиграфом $MG_f = \langle V_f, X_f \rangle$, где

$$V_f = (v_1^f, v_2^f, \dots, v_{p_f}^f) - \text{множество вершин, } p_f = |V_f|;$$

$$X_f = (x_1^f, x_2^f, \dots, x_{q_f}^f) - \text{множество дуг, } q_f = |X_f|.$$

Каждая дуга задается тройкой:

$$x_k^f = (v_i^f, v_j^f, w_k^f), \text{ где } v_i^f - \text{вершина начала дуги; } v_j^f - \text{вершина завершения дуги; } w_k^f$$

- вес дуги (идентификатор функционального отношения).

Направленный мультиграф может иметь дуги $x_k^f = (v_i^f, v_j^f, w_k^f)$ и $x_l^f = (v_i^f, v_j^f, w_l^f)$, $w_k^f \neq w_l^f$, но не может при этом содержать дугу (v_j^f, v_i^f, w)

Понятийная система определяется как:

$$S_c = \langle M_c, A_c, R_c, Z_c \rangle, \text{ где}$$

M_c - множество знаковых описаний понятий ПрО;

A_c - множество характеристических атрибутов знаковых описаний;

⁴ Так как S_f, S_c, S_t являются развивающимися системами (т.е. иметь разные состояния понятийной и знаковой систем в момент сравнения) в операциях над онтологиями они должны быть приведены к общему основанию.

⁵ Данное определение расширяет ранее упомянутое в [Van Heijst].

⁶ Закон композиции Z_f и множество характеристических атрибутов A_f представляют интенциональный способ задания онтологии. В данной статье рассматриваться не будут.

R_c – родо-видовые и ассоциативные отношения;
 Z_c – закон композиции, в соответствии с которым выбрано конкретное системное основание $\{M_c, A_c, R_c\}$.

Со структурной точки зрения понятийная система представлена помеченным взвешенным направленным графом и (с точки зрения введения операций над онтологиями) описана как

$G_c = \langle V_c, X_c \rangle$, где

$V_c = (v_1^c, v_2^c, \dots, v_{p_c}^c)$ – множество вершин, $p_c = |V_c|$;

$X_c = (x_1^c, x_2^c, \dots, x_{q_c}^c)$ – множество дуг, $q_c = |X_c|$.

Каждая дуга задается тройкой:

$x_k^c = (v_i^c, v_j^c, w_k^c)$, где v_i^c – вершина начала дуги; v_j^c – вершина завершения дуги; w_k^c –

вес дуги (идентификатор отношения).

Направленный граф не содержит симметричных дуг, любые две вершины могут быть соединены только одной дугой.

Терминологическая система:

$S_t = \langle M_t, A_t, R_t, Z_t \rangle$

M_t – множество терминов ПрО;

A_t – множество характеристических атрибутов терминов;

R_t – отношения эквивалентности и включения;

Z_t – закон композиции, в соответствии с которым выбрано конкретное системное основание $\{M_t, A_t, R_t\}$.

Со структурной точки зрения терминологическая система описывается n -связным графом $G_t = \bigcup_{i=1}^n G_i^t$, где каждая компонента связности G_i^t представляет собой полный граф (эквивалентность), дерево (включение) или результат операции объединения полных графов и деревьев (при наличии общих вершин).

Операции над онтологиями

Исследования в области онтологических представлений знаний интенсивно, хотя и достаточно эклектично развивающееся направление. Это в значительной степени относится и к введению операций над онтологиями. В многочисленных работах предлагаются достаточно разнообразные множества операций (обзоры достаточно обстоятельно представлены, например, в [FIPA 98, Proceedings, Левашова]), которые достаточно условно можно отнести к классам (1) аналоги теоретико-множественных операций (пересечение, объединение, разность); (2) операции извлечения и удаления фрагментов онтологий для использования при создании новых; (3) операции проверки логической эквивалентности онтологий, перевода онтологий на другой формальный язык и т.п.; (4) операции-функции – средства поддержки онтологий в инструментальных системах (создание, копирование, вывод и т.д.).

Рассматриваемые далее операции предназначены для построения структурных комбинаций информационных объектов (документов, сообщений), представленных онтологиями, на которых субъект может построить новое знание (задачи вывода нового знания на имеющихся онтологиях не рассматриваются).

В качестве основных операций над онтологиями (на структурном уровне) рассмотрим операции из 1-го и 2-го классов – бинарные операции объединения и пересечения и унарные – построения аспектного представления и масштабирования онтологий, с помощью которых можно, в том числе, синтезировать новые онтологии, отражающие предметную область в аспекте, заданном пользователем.

В силу того, что операции над онтологиями рассматриваются в рамках одной ПрО, будем считать, что исходные онтологии имеют общие понятийную и терминологическую системы.

Для формализации операций над онтологиями необходимо определить функцию подобия, вычисляющую меру соответствия при сравнении понятий и связей в исходных онтологиях с целью выделения семантически схожих элементов [Steffen Staab].

Введем функцию подобия $sim(a, b)$, обладающую следующими очевидными свойствами:

1. $sim(a, b) \in [0; 1]$;
2. $sim(a, b) = 1 \rightarrow a = b$ (элементы a, b тождественны);
3. $sim(a, b) = 0 \rightarrow a \neq b$ (элементы a, b различны);
4. $sim(a, a) = 1$ (свойство возвратности);
5. $sim(a, b) = sim(b, a)$ (свойство симметричности).

Функция подобия рассчитывается для элементов функциональных систем пары исходных онтологий O_1 и O_2 (в дальнейшем описанием индекс i_1 характеризует принадлежность элемента онтологии O_1 , а индекс i_2 - онтологии O_2) по следующим правилам.

Для вершин мультиграфов:

1. $sim(v_{i_1}^f, v_{i_2}^f) = 1$, если $v_{i_1}^f \equiv v_{i_2}^f$
2. $sim(v_{i_1}^f, v_{i_2}^f) = F_c(w_{k_1}^c, w_{k_2}^c, \dots, w_{k_n}^c)$, если в графе общей понятийной системы онтологий существуют принадлежащие одному маршруту вершины v_k^c и v_l^c такие, что $v_{i_1}^f \equiv v_k^c$ и $v_{i_2}^f \equiv v_l^c$ (n - длина маршрута, $w_{k_j}^c$ ($j = \overline{1, n}$) - вес отдельной дуги).

3. $sim(v_{i_1}^f, v_{i_2}^f) = f_t(p)$, если в графе общей терминологической системы онтологий существует полный подграф, содержащий вершины v_k^t и v_l^t такие, что $v_{i_1}^f \equiv v_k^t$ и $v_{i_2}^f \equiv v_l^t$ (p - количество вершин подграфа)

4. $sim(v_{i_1}^f, v_{i_2}^f) = g_t(q)$, если в графе общей терминологической системы онтологий существует цепь между вершинами v_k^t и v_l^t такими, что $v_{i_1}^f \equiv v_k^t$ и $v_{i_2}^f \equiv v_l^t$ (q - длина цепи)

5. $sim(v_{i_1}^f, v_{i_2}^f) = 0$, если не применимо ни одно из правил 1-4.

Для дуг мультиграфов:

$$sim(x_{i_1}^f, x_{i_2}^f) = \begin{cases} 1, x_{i_1}^f \equiv x_{i_2}^f \\ 0, \text{в противном случае} \end{cases}$$

Результат бинарных операций над онтологиями $O_1 = \langle S_f^1, S_c, S_t, \equiv \rangle$ и $O_2 = \langle S_f^2, S_c, S_t, \equiv \rangle$, приведенными к одному понятийному и терминологическому основанию, представляет собой онтологию $O_{op} = \langle S_f^1 op S_f^2, S_c, S_t, \equiv \rangle$. Таким образом, необходимо формализовать бинарные операции для функциональных систем.

Представление функциональных систем как мультиграфов позволяет свести операции над онтологиями к операциям над мультиграфами MG_f^1 и MG_f^2 .

При выполнении операций над мультиграфами будем считать одинаковыми вершинами те, для которых функция подобия отлична от 0, а одинаковыми дугами - те, для которых функция подобия равна 1.

Операция объединения онтологий ($O_{\cup} = O_1 \cup O_2$). Формально алгоритм объединения можно представить следующим образом:

1. Вычисляется множество вершин $V_{\cap} = V_f^1 \cap V_f^2$ (для вершин разных мультиграфов, принадлежащих множеству V_{\cap} , функция подобия равна 1).

2. Для каждой вершины из множества вершин $(V_f^1 \setminus V_\cap) \cap V_c$ формируется множество маршрутов к вершинам из множества $(V_f^2 \setminus V_\cap) \cap V_c$ в понятийном графе G_c . Если для вершины множество маршрутов не пусто, для каждого из маршрутов вычисляется функция подобия. Две вершины, для которых функция подобия принимает максимальное значение, считаются далее тождественными (в мультиграфах каждая из вершин замещается соответствующим «понятийным маршрутом») и при новом пересечении множеств вершин мультиграфов формируют множество V_\cap^c с учетом понятийной системы.

3. Рассматриваются множества вершин $((V_f^1 \setminus V_\cap) \cap V_c) \cap V_t$ и $((V_f^2 \setminus V_\cap) \cap V_c) \cap V_t$. При наличии пары вершин (по одной из каждого множества), входящих в одну компоненту связности графа терминологической системы, происходит замена каждой из вершин соответствующей компонентой (или ее фрагментом, включающим обе вершины) и при новом пересечении множеств вершин мультиграфов формируют множество V_\cap^t с учетом терминологической системы.

4. Множество вершин мультиграфа объединения онтологий формируется как результат теоретико-множественных операций:

$$V_f^\cup = V_\cap \cup V_\cap^c \cup V_\cap^t \cup (((V_f^1 \setminus V_\cap) \cap V_c) \cap V_t) \cup (((V_f^2 \setminus V_\cap) \cap V_c) \cap V_t)$$

5. Инцидентность дуг из множеств X_f^1 и X_f^2 сохраняется в мультиграфе объединения онтологий (с учетом слияния дуг, для которых функция подобия равна 1).

При объединении функциональных систем исходных онтологий по такому алгоритму могут возникнуть противоречия двух типов.

1. В результате операции объединения мультиграф онтологии O_\cup может содержать вершины из множеств V_\cap^c и V_\cap^t , которые являются подграфами понятийного и терминологического графов, дуги и ребра которых описывают отношения понятийной и терминологической систем (что противоречит определению мультиграфа функциональной системы). Для разрешения этих противоречий необходима экспертная оценка и замена таких подграфов узлами (и, возможно, дугами), соответствующими функциональной системе.

2. Сохранение в мультиграфе объединения онтологий инцидентности дуг из множеств X_f^1 и X_f^2 может привести к наличию противоречивых дуг, инцидентных одной и той же паре вершин. Такая ситуация также исследуется субъектом - экспертом.

Операция пересечения онтологий ($O_1 \cap O_2$). Отличие алгоритма пересечения от алгоритма объединения состоит в способе формирования результирующего мультиграфа:

$$V_f^\cap = V_\cap \cup V_\cap^c \cup V_\cap^t,$$

а инцидентность сохраняется в мультиграфе пересечения онтологий только для тех дуг из множеств X_f^1 и X_f^2 , для которых функция подобия равна 1.

При пересечении функциональных систем исходных онтологий могут возникнуть противоречия только первого из описанных выше типов.

Построение аспектного представления (проекция онтологии). Аспект рассмотрения (представления, описания) в свою очередь может быть задан функциональной системой $S_f^i = \langle M_f^i, A_f^i, R_f^i, Z_f^i \rangle$, тогда операция проекции может быть сведена к операции пересечения исходной - $O = \langle S_f, S_c, S_t, \equiv \rangle$ и аспектной - $O_i = \langle S_f^i, S_c, S_t, \equiv \rangle$ онтологий: $O_{proj} = O \cap O_i$. В результате операции пересечения онтологий на самом деле происходит «обогащение» аспектного описания с учетом понятийной и терминологической систем.

Рассмотрим следующие возможные ситуации.

1. $M_f^i \neq \emptyset, R_f^i = \emptyset$ - аспект задается на уровне объектов – знаковыми описаниями совокупности объектов. Мультиграф аспектной онтологии, участвующий в операции пе-

пересечения, при этом представляет собой пустой граф $MG_f^i = \langle V_f^i, \emptyset \rangle$. В этом случае для множества вершин результирующего мультиграфа V_f^{proj} операция пересечения выполняется по правилам, описанным выше, а множество дуг формируется из дуг множества X_f^i , инцидентных вершинам множества V_f^{proj} .

2. $M_f^i = \emptyset, R_f^i \neq \emptyset$ - аспект задается на функциональном уровне – множеством функциональных отношений. В этом случае множество дуг мультиграфа проекции $X_f^{proj} = X_f^i \cap X_f^i$, а множество вершин формируется из вершин, инцидентных дугам из множества X_f^{proj} .

3. $M_f^i \neq \emptyset, R_f^i \neq \emptyset$ - аспект задается на смешанном, объектно-функциональном уровне, и для операции пересечения может быть сформирован мультиграф $MG_f^i = \langle V_f^i, X_f^i \rangle$ с непустыми множествами вершин и дуг.

Масштабирование онтологий. Для описания операций масштабирования (укрупнения или детализации) онтологии определим для исходной онтологии $O = \langle S_f, S_c, S_t, \equiv \rangle$ «онтологию масштабирования» как $O_m = \langle S_f^m, S_c, S_t, \equiv \rangle$, где $S_f^m = \langle M_c, \emptyset, \emptyset, Z_f \rangle$ и сведем операции масштабирования к разрешению противоречий первого типа в онтологии $O \cup O_m$ в соответствии со следующими правилами:

- в случае укрупнения исходной онтологии: в узлах, содержащих дерево понятий, остаются понятия самого верхнего уровня, а в узлах, содержащих подграфы терминологической системы – отдельные слова;
- в случае детализации исходной онтологии: в узлах, содержащих дерево понятий, остаются понятия самого нижнего уровня, а в узлах, содержащих подграфы терминологической системы – словосочетания.

Заключение

По определению Гегеля онтология как наука - это диалектическое учение об абстрактных определениях сущности, изучающая категории философских понятий, являющиеся «кирпичиками» формирования картины мира, средством освоения человеком конкретных и абстрактных объектов.

Именно с этих позиций, в статье предложен (не претендующий на универсальность) подход к определению онтологии (в узком смысле) и множества операций как инструмента формирования и количественно оцениваемого соотношения образов объектов предметной области в условиях диалектически взаимосвязанных пространств абстрактных и конкретных объектов ПрО. В задачах информационного поиска такой подход позволит в автоматизированных процессах динамического реформулирования и соотношения поисковых образов запросов и документов учитывать разнообразие точек зрения человека – определяющего звена в процессах познания - на основе приведения поисковых образов к общему понятийно-терминологическому контексту.

Литература

1. Gruber Th. What is an Ontology // URL: <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html>
2. Van Heijst, G., Schreiber, A. T., Wielinga, B. J. 1996. Using Explicit Ontologies in KBS Development. //URL: <http://ksi.cpsc.ucalgary.ca/KAW/KAW96/borst/node16.html>
3. Guarino N. Understanding, Building, and Using Ontologies // URL: <http://ksi.cpsc.ucalgary.ca/KAW/KAW96/guarino/guarino.html>

4. Genesereth, M. R. and Nilsson, N. J. 1987. Logical Foundation of Artificial Intelligence. Morgan Kaufmann, Los Altos, California
5. David De Roure, Nicholas R. Jennings, Nigel R. Shadbolt The Semantic Grid: a Future e-Science Infrastructure // www.semanticgrid.org/documents/semgrid-journal/semgrid-journal.pdf
6. Бениаминов Е.М. Система представления знаний Ontolingua - принципы и перспективы / Бениаминов Е.М., Болдина Д. М.// НТИ, сер. 2, 1999, №10
7. Мудрая О.В. , Бабич Б.В. , Пьяо С.С. , Рейсон П. , Уилсон Э. Разработка инструментария для семантической разметки текст // Труды международной конференции "Корпусная лингвистика–2006". Издательства СПбГУ и РХГА, 2006.
8. *M. Uschold, M. King, S. Moralee, Y.Zorgios.* The Enterprise Ontology. -The Knowledge Engineering Review, v.13, N1, 1998, p.p. 31-88.
9. Урманцев Ю.А. Общая теория систем: Состояние, приложения и перспективы развития// Сб. «Система, Симметрия, Гармония». – М.: Мысль, 1988. – С. 38-124.
10. Максимов Н.В. Информация и знания: природа, концептуальная модель // НТИ сер.2, 2010, №7, с.1-10.
11. Бониц М. Научное исследование и научная информация. –М.: Наука, 1987, 156с.
12. 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. Серия «Управление», № 4 (66) / 2011. –М.:РГГУ, 2011 – С. 175-185
13. Nirenburg S., Raskin V.. Ontological Semantics. Cambridge, MA, 2004
14. FIPA 98 Specification. Part 12. Ontology Service. — Geneva, Switzerland: Foundation for Intelligent Physical Agents (FIPA), 1998. — Version 1.0.
15. Proceedings of the Workshop on Application Integration Architectures (eds. by Hodges B.,Thompson C.). — Texas Instruments, NIST, 1993.
16. Левашова Т. В. Принципы управления онтологиями, используемые в среде интеграции знаний // Труды СПИИРАН. Вып. 1, т. 2. — СПб.: СПИИРАН, 2002.
17. Steffen Staab, Rudi Studer, Prasenjit Mitra, Gio Wiederhold «Handbook on ontologies» //Springer, Germany, Berlin, 2004

Ключевые слова

Онтологии; операции над онтологиями; информационный поиск; идентификация содержания; общая теория систем; теория графов.

Сведения об авторах:

Голицына Ольга Леонидовна – доц., к.т.н., доцент каф. Системного анализа Национального исследовательского ядерного университета «МИФИ»

тел. 8-906-095-71-21

eMail: OLGolitsina@YANDEX.RU

Максимов Николай Вениаминович – проф., д.т.н., профессор каф. Системного анализа Национального исследовательского ядерного университета «МИФИ»

тел. 8-903-683-12-83

eMail: NV-MAKS@YANDEX.RU

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

тел. 8-916-267-42-08

eMail: doomguard@YANDEX.RU