

Особенности использования языков семантических запросов и правил Semantic Web в интеллектуальных системах, основанных на технологиях концептуальных графов

С. В. Власенко

Санкт-Петербургский государственный электротехнический университет

«ЛЭТИ» им. В.И. Ульянова (Ленина)

s_v_vlasenko@mail.ru

Аннотация. В статье рассматриваются проблемы применения языков семантических запросов и правил, ориентированных на использование в рамках Semantic Web, в интеллектуальных информационных системах, основанных на аппарате концептуальных графов и моделях знаний соответствующего вида. Особое внимание при этом уделяется анализу корректности интерпретаций данных языков, а также технологическим аспектам организации обработки моделей знаний выбранного класса.

Ключевые слова: интеллектуальные системы, модели знаний, Semantic Web, концептуальные графы

I. ВВЕДЕНИЕ

Развитие концепций «цифровой экономики», Industry 4.0, технологий Big Data, «Интернета вещей» и многие другие тенденции в современной техносфере постоянно стимулируют динамичные изменения в областях науки и технологий, связанных с проблематикой распределенной семантической обработки данных и знаний на основе М2М-взаимодействий и автоматического обеспечения интероперабельности информационных систем и ресурсов. Значительное разнообразие идей и подходов к решению соответствующих технических задач привело к появлению и параллельному эволюционному развитию большого числа международных, национальных и корпоративных исследовательских программ, проектов, технологий, стандартов и рекомендаций, программных инструментов и т.д., применяемых в рассматриваемой предметной области. При этом неизбежно возникает и постепенно расширяется круг проблем прикладного уровня, обусловленных, по большей мере, отсутствием согласованности различных используемых на практике технологических инноваций и/или теоретических моделей. Показательным примером подобной ситуации является положение дел с пулом технологий, развиваемых в рамках программы корпорации W3C Semantic Web, и с альтернативными технологиями семантической обработки данных и знаний, в том числе – технологиями, базирующимися на стандартизированных в ISO моделях, методах и инструментах.

Безусловно, в рамках данной статьи невозможно охватить весь спектр аспектов и отдельных вопросов, которые подлежат рассмотрению в контексте заявленной темы. Конкретным предметом анализа в настоящей статье является только одна из «составляющих» описываемой комплексной проблемы – соотнесение задач применения регламентированных в Semantic Web аппаратов семантических запросов и правил с задачами обработки моделей, построенных на базе аппарата концептуальных графов (КГ) ([1–3]), являющегося одним из компонентов стандарта ISO «Common Logic» ([4]). При этом непосредственные постановки рассматриваемых вопросов связаны, прежде всего, с задачами инициативного проекта по разработке технологии и комплекса программных средств поддержки проектирования интеллектуальных систем AIS ([5]), разрабатываемого на кафедре Автоматики и процессов управления Санкт-Петербургского государственного электротехнического университета «ЛЭТИ». Одним из компонентов перспективной версии данного комплекса (в [5] – «AIS-2») является виртуальная машина КГ (CGVM), в функции которой планируется включить обработку внешних семантических запросов и правил, представленных в наиболее распространенных нотациях, включая языки, предлагаемые в рамках Semantic Web.

II. COMMON LOGIC И SEMANTIC WEB

Исходные цели создания подходов и инструментов, регламентируемых в Common Logic (CL) и в Semantic Web, в значительной мере совпадают, хотя официально CL декларируется лишь как стандарт на семейство логических языков, предназначенных для представления и обмена информацией между независимыми информационными системами, а Semantic Web – как совокупность стандартов W3C, направленных на обеспечение «машинной читаемости» данных в сети Internet (что является гораздо более обширной технологической проблемой). В практическом плане основным элементом «пересечения интересов» для двух названных технологических инициатив являются выразительные средства, используемые в процессах представления и обработки

данных и знаний, а также лежащие в их основе формальные логические и информационные модели. В конкретном выражении данные пересечения охватывают, прежде всего, следующие компоненты рассматриваемых стандартов:

в Semantic Web –

- RDF (Resource Description Framework);
- RDFS (RDF Schema);
- OWL (Web Ontology Language);
- SPARQL (SPARQL Protocol and RDF Query Language);
- RIF (Rule Interchange Format);

в Common Logic –

- CLIF (Common Logic Interchange Format);
- CGIF (Conceptual Graph Interchange Format).

Учитывая тематический профиль настоящей статьи, остановимся далее на анализе трех компонентов – нотации CGIF (включая диалект Extended CGIF) в CL (как наиболее выразительного инструмента CL в задачах представления и обработки моделей знаний и данных общего вида) и нотаций SPARQL ([6]) и RIF ([7]) в Semantic Web. При этом основное внимание будет уделено трем вопросам:

1) могут ли в принципе SPARQL и RIF корректно (с точки зрения логических интерпретаций) использоваться применительно к моделям знаний, представленным в виде концептуальных графов (КГ), построенных на базе нотаций CGIF;

2) имеет ли поддержка отдельных функциональных возможностей SPARQL и RIF практический смысл в контексте их применения к КГ-моделям;

3) каковы области применимости и ограничения при использовании SPARQL и RIF в соответствующих приложениях.

Безусловно, на уровне общетеоретических положений ответы на сформулированные выше вопросы в значительной мере определяются общей «иерархией» формальных моделей, предлагаемых стандартами CL и Semantic Web. Данная «иерархия» может быть условно установлена на основании того факта, что языки RDF/RDFS и OWL (а, соответственно, и ориентированные на совместное с ними использование нотации SPARQL и RIF) являются семантическими подмножествами CL и базовых диалектов CL – CLIF и CGIF. Однако на уровне практических приложений речь, как правило, идет о специфике «реализаций» аппаратов CL и Semantic Web в конкретных проектах и решениях. Именно на этом и будет сфокусирован последующий материал настоящей статьи. При этом практические аспекты соответствующих проблем будут рассматриваться применительно к контексту задач, связанных с созданием комплекса AIS-2, о котором говорилось ранее в разделе I.

III. ОСЛУЖИВАНИЕ SPARQL-ЗАПРОСОВ

Поддержка SPARQL в виртуальной машине КГ (CGVM) имеет достаточно важное значение, поскольку позволяет включать модели на базе КГ в интегрированные решения по распределенной обработке знаний и данных наряду, например, с OWL-онтологиями, базами данных и другими традиционными видами источников информации для SPARQL-точек.

Технологически обработка SPARQL-запроса в CGVM в общем виде может быть описана как последовательность выполнения следующих этапов:

(I) декомпозиция запроса с формированием множества конструктивных элементов, описываемых RDF-графами;

(II) трансляция элементов запроса в выбранную CL-нотацию (в данном случае – в CGIF);

(III) вычисление результата на базе семантики CGIF;

(IV) приведение полученного результата к RDF-форме, соответствующей параметрам запроса.

Разумеется, конкретизация действий на каждом из перечисленных обобщенных этапов зависит от формы запроса (SELECT, CONSTRUCT, ASK или DESCRIBE), параметров фильтрации, модификаторов результата и т. д. Однако уже на уровне обобщенного алгоритма обработки запроса должно учитываться следующее обстоятельство: если на этапах (I)-(III) с учетом соблюдения принятых в стандарте SPARQL 1.1 ограничений на формы шаблонов RDF-триплетов может быть гарантирована корректность соответствующих преобразований, то на этапе (IV) данные гарантии не обеспечиваются. Между тем, КГ-модель, к которой осуществляется запрос, может содержать вполне состоятельный ответ на него, и этот ответ может иметь практическую ценность даже в случае неоднозначности или возможности расширенной трактовки результата.

Для разрешения указанной конфликтной ситуации в рамках проекта CGVM планируется применять подход, основанный на структурных особенностях КГ-моделей на базе нотаций Extended CGIF, и затрагивающий этапы (II), (III) и (IV) в приведенном выше обобщенном алгоритме.

Наличие формализованных соответствий нотаций и структур моделей в RDF/RDFS, OWL и КГ значительно упрощает процессы анализа корректности семантики запросов и содержания их интерпретаций. Очевидно, что на структурном уровне принципиальными барьерами для прямого сопоставления КГ с множествами RDF-триплетов «субъект – предикат – объект» (задающих «семантический каркас» для получения результатов SPARQL-запроса) являются:

- доступное в аппарате КГ рекурсивное определение концептов («контекстов», т. е., концептов, содержащих вложенные в них КГ-подграфы);
- наличие класса кореферентных ссылок, задающих отношения эквивалентности концептов в различных контекстах;

- различия в номенклатуре и способах структурного отображения предопределенных контейнеров;
- возможное наличие в КГ-модели предопределенных специализированных контекстов (в том числе – логических операций) и акторов (как особого класса концептуальных отношений).

В рамках предлагаемого подхода к построению метода обработки SPARQL-запросов на этапе трансляции RDF->CGIF (этап II) концепты и отношения в результирующих КГ-графах типизируются в соответствии с нормами CGIF и с учетом возможной необходимости адаптации описаний типовых контейнеров в исходной формулировке запроса.

На этапе (III) («вычисление результата») применяется схема «послойной» рекурсивной обработки КГ-источника с поиском проекций по каждому из элементов запроса. Основная идея данного алгоритма сводится к следующему (на примере обработки «SELECT»-запроса).

КГ-источник данных рассматривается как множество КГ нулевого уровня вложенности. Для каждого j -го КГ на i -ом уровне вложенности справедливо: $cg(i,j) = \{C(i,j), R(i,j), E(i,j)\}$; $cg(i+1,j) \subset C(i,j)$, где C и R – множества концептов и отношений в $cg(i,j)$, а E – множество дуг данного графа, отвечающих условиям двудольности графа по подмножествам вершин C и R . Тогда обработка $cg(i,j)$ при анализе произвольного k -го элемента запроса $cgx(k)$, где $cgx(k) = \{Cx(k), Rx(k), Ex(k)\}$ – КГ данного элемента, не имеющий вложенных контекстов и соответствующий RDF-шаблону принятой в запросе структуры, сводится к последовательному поиску проекций в соответствующем контексте КГ-источника с выполнением следующей последовательности действий (назовем ее процедурой $S(i,j,k)$):

- исключение из дальнейшего процесса обработки всех специализированных контекстов логических операций и семантических правил (например, контекстов «if»/«iff» или отрицаний), не отвечающих элементам запроса, заданным в предложениях «WHERE», а также контекстов, которые подпадают под условия фильтрации в соответствии с содержанием предложений «FILTER»;
- поиск соответствий (с учетом фактуры шаблона) между $cgx(k)$ и множеством триплетов $C(i,j) \times R(i,j) \times C(i,j)$, в котором каждая упорядоченная пара $\langle c,r \rangle$ из узла-концепта и узла-отношения представлена в $E(i,j)$. При выявлении соответствия найденный триплет дописывается (с сохранением и унификацией вмещающего контекста) в текущий граф результата (по k -му маршруту);
- разрешение кореферентных ссылок: для потенциально подходящих под шаблон $cgx(k)$ КГ-триплетов при наличии в соответствующих концептах $cg(i,j)$ меток «исходящих» кореферентных связей (меток вида [... *x]) процедура $S(i,j,k)$ последовательно выполняется для

всех вложенных в $cg(i,j)$ контекстов, содержащих «входящие» метки этих связей (вида [... ?x]). При этом в случае нахождения корректных проекций $cgx(k)$ на базе дизъюнктивных CGIF-контекстов могут формироваться альтернативные ветки k -го маршрута в графе результата;

- рекурсивный запуск процедуры $S(i,j,k)$ для всех значений j , определенных в $cg(i,j)$.

Обработка промежуточных результатов выполнения запроса зависима от выбираемого способа обслуживания КГ-ответа на каждой итерации реализации процедуры $S(i,j,k)$. Различные сценарии построения КГ-ответа (графа cgr) в целом сводимы к двум базовым схемам:

- с независимым построением $cgr(k)$ по каждому k -му элементу запросу и с последующей композицией частных результатов;
- с динамическим наращиванием cgr по результатам исполнения каждого экземпляра процедуры $S(i,j,k)$.

Вне зависимости от выбранной схемы непосредственно процесс композиции КГ при построении cgr базируется на оговоренных в рамках CL и в теории КГ правилах логического вывода и формальных преобразований КГ (обобщениях, конкретизациях, объединениях, упрощениях и т. п.).

При реализации IV этапа обработки SPARQL-запроса (формирование выходного представления результатов) полученный в ходе «послойного» обслуживания КГ-ответ (результата выполнения этапа III) позволяет синтезировать итоговый RDF-граф простейшими способами, поскольку к этому моменту все основные структурные особенности организации КГ, отмечавшиеся ранее в данном разделе, оказываются учтенными за счет декомпозиции иерархий вложенности контекстов, раскрытия кореферентных связей, вынесения решений задач динамической типизации узлов за пределы поисковой процедуры и т. д.

Детализированное изложение рассмотренных методов обслуживания SPARQL-запросов на базе аппарата КГ и нотаций CGIF, вероятно, выходит за тематические рамки настоящей статьи. Однако представляется, что описанные основные элементы предлагаемых решений позволяют сделать общие выводы о предметном содержании описанного выше подхода. В перспективе развитие этого метода и программная реализация соответствующих функций в проекте CGVM может привести к возможности поддержки решения обширного класса прикладных задач интеграции разнородных моделей знаний и данных с применением распространенных языков Semantic Web. Кроме того, будут обеспечены необходимые предпосылки и для решения обратных задач, связанных, например, с автоматической генерацией в среде интеллектуальных систем, основанных на аппарате КГ, SPARQL-текстов для реализации процедур семантического поиска и других процессов приобретения и обработки знаний, доступных во внешних информационных источниках.

IV. ПРИМЕНЕНИЕ RIF

Задачи обеспечения поддержки языков семантических правил семейства RIF применительно к отмеченным ранее целям построения функционально развитой CGVM во многом пересекаются с рассмотренной в разделе III проблемой реализации функций SPARQL-точки в среде AIS-2. Общим моментом здесь является привлечение регламентируемых в Semantic Web средств работы с моделями знаний и данных, базирующихся на стандартах RDF/RDFS и OWL, к процессам обработки информации в ориентированных на использование КГ интеллектуальных системах.

На текущем этапе развития проекта AIS-2 планируется разработка средств поддержки трех базовых диалектов RIF:

- RIF Core (диалект ядра RIF);
- RIF-BLD (BLD – Basic Logic Dialect – основной логический диалект);
- RIF-PRD (PRD – Production Rule Dialect – диалект производственных правил).

В приведенном выше перечне RIF Core является общим семантическим подмножеством для RIF-BLD и RIF-PRD, RIF-BLD представляет собой базовый диалект в классе логических («logic-based») диалектов RIF (к числу которых относится и RIF Core), а RIF-PRD является базовым диалектом в классе «rules-with-actions», т. е. ориентирован на обработку «исполняемых» продуктов.

В контексте настоящей статьи наиболее важным общим свойством всех RIF-диалектов (включая перечисленные ранее) является регламентированная совместимость со стандартами RDF, OWL и SPARQL (3 из 12 документов, составляющих стандарт RIF, посвящены именно этим вопросам). Данную совместимость обеспечивает наличие семантически корректных взаимных преобразований и интерпретаций RIF-нотаций и RDF-триплетов (в том числе – на уровне RDFS- и OWL-онтологий). Соответственно, для случая обслуживания на базе средств RIF КГ-моделей, представимых в Extended CGIF, могут использоваться все рассмотренные выше в разделе III механизмы совместной обработки RDF- и КГ-графов.

В рамках разработки проекта AIS-2 и виртуальной машины CGVM основными сценариями применения RIF на данном этапе будут являться:

- поддержка представленных в виде RIF-нотаций (BLD и Core) логических выводов на КГ-моделях;
- управление КГ-моделями знаний и данных на базе RIF-правил;
- семантический поиск знаний и данных в КГ-моделях на основании логических заключений;
- интеграция CGVM в BRMS-системы (Business Rule Management System) с совместным использованием КГ-ориентированных баз знаний и RIF-PRD;
- поддержка разработки распределенных экспертных систем производственного типа на базе КГ-моделей знаний.

В перспективе использование RIF в проекте AIS-2, безусловно, позволит поддерживать и решение значительно более широкого класса задач обработки знаний и данных в веб-ориентированных приложениях различного уровня и функционального профиля.

V. ВЫВОДЫ

Кратко рассмотренные в данной статье подходы к обеспечению возможности использования ряда базовых инструментов Semantic Web в технологиях и системах, построенных на применении аппарата КГ, разумеется, требуют значительной доработки и развития по целому ряду позиций. Тем не менее, представляется, что решение обсуждавшихся выше вопросов потенциально может иметь достаточно весомое практическое значение. В сфере развития технологий Semantic Web растет актуальность задач существенного повышения выразительной мощности средств, применяемых для организации информационных взаимодействий в М2М-средах, особенно в классе средств представления знаний. С другой стороны, обладающие значительной семантической мощностью и потенциалом расширения функциональных возможностей инструменты, предоставляемые аппаратом CL, на данный момент очень слабо поддержаны существующими технологическими и программными решениями, включая инфраструктурный уровень Интернет-пространства. Из сказанного следует, что взаимное «сближение» двух соответствующих стеков технологий, вероятно, могло быть полезно для обеих технологических инициатив. Однако на данный момент в материалах рабочих групп W3C и ISO/IEC, занимающихся соответствующими разработками, а также в действующих стандартах и рекомендациях тенденции к сближению, о котором идет речь, не прослеживаются. Возможно, что появление и развитие прикладных проектов и решений, интегрирующих элементы технологий Semantic Web и CL (к числу которых относится и рассматриваемый в статье проект AIS-2), позволит «сократить дистанцию» между соответствующими программами исследований и создаст определенные предпосылки для формирования в будущем более взаимосвязанных регламентов в стандартах W3C и ISO.

СПИСОК ЛИТЕРАТУРЫ

- [1] John F. Sowa. Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA USA, 2000. 594 p.
- [2] Michel Chein, Marie-Laure Mugnier: Graph-based Knowledge Representation - Computational Foundations of Conceptual Graphs. Advanced Information and Knowledge Processing, Springer 2009, ISBN 978-1-84800-285-2, pp. 1-427.
- [3] Handbook of Knowledge Representation / Ed. by Frank van Harmelen, Vladimir Lifschitz, Bruce Porter. Elsevier, 2008. 1034 p.
- [4] ISO/IEC 24707:2018 Information technology – Common Logic (CL): a framework for a family of logic-based languages. International Organisation for Standardisation (Second Edition).
- [5] Vlasenko S., Efimenko G., Gnezdilov D., Brikova O. Approaches to Conceptual Graphs Notations Using in Digital Manufacturing Software Environments // Proceedings. 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (2019 ElConRus) p. 731-735. 10.1109/ElConRus.2019.8656842.
- [6] SPARQL 1.1 Query Language. W3C Recommendation, 2013. <http://www.w3.org/TR/2013/REC-sparql11-query-20130321/>.
- [7] RIF Overview (Second Edition). W3C Working Group Note, 2013. <http://www.w3.org/TR/2013/NOTE-rif-overview-20130205/>