МОДЕЛЬ МЕТАДАННЫХ ДЛЯ СЕМАНТИЧЕСКОГО ПОИСКА РЕАЛИЗАЦИЙ ПОТОКОВ РАБОТ, ВЫРАЖЕННЫХ В ВИДЕ ПРАВИЛ*

Н. А. СКВОРЦОВ[1]

А. Е. ВОВЧЕНКО[2]

Л. А. КАЛИНИЧЕНКО[3]

Д. А. КОВАЛЕВ[4]

С. А. СТУПНИКОВ[5]

Аннотация: Конструирование потоков работ и накопление доступных методов научными сообществами предполагает спецификацию их в открытой среде и организацию их поиска с целью использования при решении задач. В данной работе для спецификации потоков работ используются диалекты языка правил RIF, а для организации поиска в предметной области элементы потоков работ аннотируются метаданными. Модель метаданных, необходимых для семантического поиска, включает описание структуры скелета потока работ, связывание элементов потоков работ с понятиями предметной области, требования к качеству и происхождению данных и методов. Метаданные определяются в терминах соответствующих онтологий. Пример демонстрирует поиск реализаций потоков работ в коллекции методов для составления потока работ из существующих релевантных компонентов.

* Работа выполнена при поддержке РФФИ (гранты 13-07-00579, 14-07-00548), Президиума РАН (Программа фундаментальных исследований Президиума РАН, тема 16-3 «Фундаментальные проблемы системного программирования»), ИПИ РАН (Тема 38.25 «Спецификация и решение задач анализа данных в концептуальных терминах предметных областей с интенсивным использованием данных» государственного задания ФГБУН ИПИ РАН).

[1] Институт проблем информатики РАН, nskv@ipi.ac.ru

[2] Институт проблем информатики РАН, itsnein@gmail.com

[3] Институт проблем информатики РАН, leonidk@ipi.ac.ru

[4] Институт проблем информатики РАН, dm.kovalev@gmail.ru

[5] Институт проблем информатики РАН, ssa@ipi.ac.ru

Ключевые слова: потоки работ на правилах, метаданные, семантический поиск

Keywords: rule-based workflows, metadata, semantic search

1 Введение

В сообществах, представляющих различные направления науки, разрабатывается и применяется собственный ассортимент методов, включающий оценку существенных свойств и параметров объектов, выявление особенных объектов, нахождение фактов, подтверждающих или опровергающих экспериментальные модели и гипотезы. Большая часть этих методов достаточно постоянна для определённой предметной области. Сырые данные подвергаются автоматической обработке некоторым набором методов, а затем результаты их применения используются для специализированных исследований, также использующих свои наборы методов. Разрабатывать все применяемые методы каждый раз при решении новой научной задачи нет смысла, если в экспертном сообществе предметной области присутствуют доступные и подходящие реализации необходимых методов. Поэтому обработка больших объёмов данных и расширение направлений их обработки при научных исследованиях заставляет подходить к множеству средств обработки данных как к коллекциям научных методов.

Для организации обработки данных целесообразно разрабатывать потоки работ, которые представляют собой спецификации порядка обработки данных и используют существующие ресурсы, сервисы, другие потоки работ или их фрагменты. Спецификации потоков работ определяют, какие входные данные необходимы для работы методов, что и по каким алгоритмам они вычисляют, в какой последовательности применяются и какие результаты выдают. Подход к обработке данных и решению задач над данными с конструированием потоков работ удобен как для составления наборов методов

для автоматической обработки данных, так и их применения для решения задач над данными.

В данной статье, являющейся расширением и дополнением исследования [1], для спецификации потоков работ используются языки и технологии, применяемые в рамках Семантического веба. Основным средством спецификации потоков работ являются диалекты языка правил RIF [2]. Концептуальные схемы предметных областей, над которыми разрабатываются спецификации деятельностей, описываются средствами языка онтологий OWL 2.0 [3] и импортируются в спецификации потоков работ [4].

Поиск реализаций потоков работ, соответствующих спецификациям, производится среди описаний потоков работ и их фрагментов в доступных коллекциях научных методов. Для возможности семантического поиска в таких коллекциях реализации потоков работ, помимо спецификации их структуры, сопровождаются определённым набором метаданных, несущих информацию о связи потоков работ с понятиями предметной области, о качестве и происхождении используемых данных и методов [5]. Эта информация обеспечивает не только возможность оценки потоков работ и их фрагментов с точки зрения структуры, но и учёт семантики предметной области и требований к качеству и надёжности работы научных методов.

Принципы семантического поиска подходящих реализаций потоков работ и их фрагментов на основе метаданных, а также модель необходимых метаданных являются предметом исследования данной статьи. В следующем разделе приведён обзор проектов, разрабатывающих средства накопления потоков работ и методы их поиска в коллекциях. Затем в разделе 3 определяются требования к спецификациям потоков работ для возможности поиска и повторного использования адекватных реализаций при решении задач, и описывается модель потоков работ на правилах и принципы связывания метаданных со спецификациями потоков работ. В разделе 4 описана модель метаданных, используемая в исследовании. Описание подхода к поиску потоков работ по метаданным приведено в разделе 5. В разделе 6

рассматривается пример сценария семантического поиска релевантных потоков работ по метаданным.

2. Родственные работы

Коллекции научных методов и потоки работ разрабатываются и занимают своё место в инструментарии научного сообщества, в первую очередь, в науках с интенсивным использованием данных.

Так, в астрономии работают программные сервисы, в которых есть возможность реализовать потоки работ, сохранять их в виде скриптов на внутренних языках систем. По такому принципу организованы онлайн-сервис Aladin [6], предоставляющий средства визуализации данных Страсбургского центра данных; свободно распространяемый продукт Topcat [7], предназначенный для работы с таблицами и включающий средства обработки таблиц, создания подмножеств данных, ведения статистики, визуализации и прочие. Сеть AstroGrid [8] представляет собой инфраструктуру для решения задач на множестве узлов, предоставляющих всевозможные сервисы и ресурсы, которые можно найти на основе метаданных через общий реестр. Имеются средства разработки приложений над доступными сервисами. На сегодня работы по развитию проекта закрыты, в первую очередь, по причине медленного развития сети, организация узлов сети оказывается сложна для широкого распространения в астрономическом сообществе. Описанные инструменты не имеют средств поиска потоков работ. В Astrogrid разработанные приложения могут быть описаны метаданными и опубликованы в реестре в качестве сервисов.

Распространенным средством разработки потоков работ стал инструмент Taverna [9], разработанный группой myGrid, в первую очередь, для сообщества исследователей в области биоинформатики, но используемых и другими сообществами исследователей. Потоки работ здесь конструируются в простой модели с использованием блоков из библиотек сервисов.

Помимо общеупотребимых сервисов разрабатываются специализированные библиотеки для определённых научных областей. В частности, накапливаются библиотеки сервисов в биоинформатике, астрономии, мультимедиа и других областях. Коллекция научных потоков работ MyExperiment [10], среда поддержки которой также разработана группой myGrid, объединяет тысячи пользователей и потоков работ и десятки проектов, предоставляющих и использующих накопленные потоки работ, в основном, разработанные в Taverna. Среда MyExperiment организована как социальная сеть, позволяющая регистрировать исследователей, включать их в различные тематические группы, публиковать потоки работ, реализованные в различных сторонних системах, описывать эксперименты, связанные с вызовом потоков работ, составлять объекты исследования, состоящие из потоков работ, документов, файлов данных. Среда MyExperiment обеспечивает поиск потоков работ по метаданным, предоставляет их описание, позволяет их запускать. Однако у данной среды есть ряд недостатков, препятствующих повторному использованию потоков работ. Во-первых, потоки работ могут специфицироваться в различных внешних редакторах и форматах. Нет общего языка спецификаций потоков работ, реализуемых в разных системах, внутренняя структура потоков работ не имеет интерфейса доступа. Во-вторых, многие потоки работ в MyExperiment обращаются к конкретным источникам данных и не имеют возможности подмены источников. В коллекции есть разработки, которые по своей сути являются не реализациями методов, а сервисами, предоставляющими данные из специфических источников данных по некоторым входным параметрам. В-третьих, для обеспечения поиска потоков работ в среде поддерживаются только вербальные пояснения к потокам работ в целом и теги.

В Taverna поддерживаются спецификации происхождения данных. Однако предназначены метаданные о происхождении только для записи пути прохождения данных внутри исполненного потока работ. Достоверность данных источника данных и реализаций методов для использования в решении

задач невозможно отследить без истории их получения и преобразования от момента создания. К тому же нет доступа через интерфейсы MyExperiment к информации о пути прохождения данных в потоке работ.

С системами Taverna и MyExperiment связаны проекты, предназначенные для поддержки жизненного цикла потоков работ и дополняющие их функциональность. Проект wf4ever [11] предоставляет набор средств для спецификации происхождения данных, внутренней структуры потоков работ, поддержки их долговременного хранения, развития, многоверсионности, проверки на доступность и совместимость источников данных. Для этого предоставляются необходимые структуры данных и интерфейсы пользователя. Спецификации предметов исследования и потоков работ можно импортировать из MyExperiment, дополнять спецификациями, предоставляемыми проектом, и использовать набор сервисов для поддержки жизненного цикла потоков работ. Проект не предполагает продвижения в сторону семантических подходов к обеспечению доступа к потокам работ, а направлен больше на разработку сервисов анализ самих потоков. Операции жизненного цикла сервисов и потоков работ, разработанные группой myGrid, включают реализацию и идентификацию повторно используемых сервисов и потоков работ, регистрацию доступных сервисов и потоков работ для использования в сообществах, аннотирование зарегистрированных спецификаций, поиск среди доступных сервисов и потоков работ потребителями, повторное использование обнаруженных сервисов и потоков работ, ведение истории их повторного использования [12].

Одной из ключевых функциональных возможностей сред поддержки научных исследований, накапливающих сервисы и потоки работ, является их поиск. Подходы к поиску потоков работ делятся, по существу, на два основных направления:

- поиск по метаданным, аннотирующим потоки работ;
- оценка близости структур потоков работ.

В целом, в большинстве проектов, посвящённых поиску потоков работ и принадлежащих первому направлению, состав метаданных ограничивается набором предопределённых свойств для работы с простыми сопроводительными данными: имёнами, вербальными определениями, информацией об авторах, версиях, правах, дате создания и другими достаточно ограниченными описаниями [13]. Такие подходы к спецификации метаданных представляются недостаточными для выразительного семантического описания и поиска потоков работ.

К первому направлению также относятся исследования поиска потоков работ в разработках myGrid [12]. При регистрации в репозитории myGrid потоки работ аннотируются описаниями четырёх уровней:

- вербальным описанием;
- описанием интерфейса потока работ как сервиса в терминах контролируемого словаря;
- описанием компонентов потока работ в терминах контролируемого словаря;
- описанием операций, производимых компонентами.

Эти описания позволяют производить поэтапный поиск потоков работ, от оценки релевантности по вербальному описанию до проверки функциональности элементов найденных потоков работ.

Проекты, посвященные оценке близости потоков работ по их графовой структуре, часто используют эвристические методы. Целый класс проектов основан на разновидностях подходов CBR (Case-Based Reasoning) [14], основанных на понятиях близости и расстояния между наборами значений свойств объектов, используемых соответственно для оценки близости объектов и их адаптации к запросу. В простых случаях они не учитывают графовую структуру потоков работ [15]. В большинстве работ при определённой модели графовых структур потоков работ сравниваются только текстовые метки [16]. В [17] потоки работ представлены в виде аннотированных графов, и разработана

модель вычисления меры близости, учитывающая как структуру потоков работ, так и связанные с ней текстовые описания.

Ещё одно распространённое направление - process mining [18] - специализируется на анализе лог-файлов. В исследованиях используются модели процессов, являющиеся спецификациями структуры потоков работ. Записи логов исполняемых деятельностей или происходящих событий сопоставляются со спецификациями моделей процессов. На основе лог-файлов решаются следующие виды задач:

- *задача обнаружения* (discovery) потоков работ – восстановление фактической структуры потока работ по лог-файлам работы его экземпляра. Таким образом, могут быть обнаружены потоки работ, не имеющие формальных спецификаций модели процесса;
- *задача установления конформности* (conformance) потока работ заключается в проверке соответствия модели потока работ данным, получаемым из лог-файлов о работе его реализации;
- *задача усовершенствования* (enhancement) модели потока работ отличается от задачи установления конформности тем, что модель не только оценивается на соответствие реальным событиям, но и меняется для более точного соответствия.

Перечисленные подходы полезны для поиска потоков работ по спецификации их структуры, для описания и дальнейшего повторного использования доступных потоков работ, не имеющих формальных спецификаций, но генерирующих лог-файлы во время своей работы, для контроля соответствия реализованных и найденных потоков работ спецификациям.

3. Требования к спецификациям потоков работ и используемая модель спецификаций

Для создания инфраструктур поддержки научных исследований на основе доступных сервисов и потоков работ информационные ресурсы и реализации научных методов должны изначально разрабатываться с учётом такой возможности, систематизироваться и описываться семантически в терминах предметной области. Это позволяет упростить интеграцию информационных и методических ресурсов. Реализации научных методов необходимо разрабатывать таким образом, чтобы упростить или даже автоматизировать их семантический поиск и использование в соответствии со спецификациями предметной области, развивать инфраструктуры предметных областей и собирать в них коллекции научных данных и методов. В этой связи в основе разработки ресурсов научных данных, методов и потоков работ должны лежать определённые принципы.

Для развития семантических подходов к решению научных задач данные, информационные ресурсы и реализации научных методов необходимо связывать со спецификациями предметной области. Под спецификацией предметной области, доступной и принимаемой сообществом исследователей, можно понимать набор связанных формальных онтологий предметной области исследования и смежных с ней областей. В соответствии с онтологиями могут создаваться концептуальные схемы предметной области, необходимые для организации информационных структур и спецификации методов, используемых в обработке данных. Спецификации методов и потоков работ должны быть декларативно определены в предметной области. Описание структуры сервисов и потоков работ должно содержать подробные спецификации пред- и пост-условий их выполнения. Агентами научного сообщества могут выступать как исследователи, так и информационные системы. Поэтому спецификации, описывающие методы и данные, должны обеспечивать понимание человеком и возможность машинной обработки.

Научные методы и данные должны быть доступны для использования научным сообществом, работающим и решающим задачи в данной предметной области. Для этого они должны быть надлежащим образом специфицированы и

опубликованы в общедоступных коллекциях. Коллекции обеспечиваются средствами семантического поиска.

Важными принципами составления спецификаций научных методов и потоков работ является их абстрактность по отношению к реализациям и независимость от конкретных источников данных. Подмена источников данных другими релевантными источниками надлежащего качества должна быть проста и не должна сказываться на работоспособности методов.

При повторном использовании реализаций методов и потоков работ необходимо быть уверенными в надёжности их работы и достоверности используемых данных. Для этого и данные, и методы необходимо сопровождать информацией об их происхождении. Она включает аутентификацию, источники, версии, трансформации от создания до момента использования. С другой стороны, реализации методов, либо системы, в которых они реализуются, должны сохранять информацию о происхождении обрабатываемых данных и обеспечивать дополнение этой информации в соответствии с манипуляциями, производимыми ими над данными. Достоверность данных и надёжность методов также определяется их качеством: точностью, полнотой, актуальностью открытых используемых данных и получаемых результатов, обеспечиваемых научными методами.

Для обеспечения повторного использования методов и потоков работ спецификации могут включать требования к средам реализации спецификаций, с указанием поддерживаемых стандартов, протоколов, моделей данных. Для проверки правильности реализации функционирование сервисов и потоков работ может проверяться наборами тестов, сопровождающими спецификации и учитывающими максимальное количество особых случаев.

Предъявленные требования к спецификациям выполняются с использованием модели, используемой для спецификации потоков работ в данной работе. Модель потоков работ на правилах в диалектах языка RIF обеспечивает соответствие требованиям декларативности и глубины формальной спецификации, абстрактности спецификаций, требованиям

независимости от конкретных информационных источников. Использование средств Семантического Веба обеспечивает понимание человеком и автоматическую обработку спецификаций.

Схема данных определяется в модели языка OWL 2.0. Язык правил используется для спецификации действий, производимых над схемой данных при решения задач, и для оркестровки потоков работ, определяющей условия и последовательность выполняемых действий.

Потоки работ специфицируются в мультидиалектной среде [4]. Определенные в концептуальных схемах сущности могут использоваться в качестве предикатов в правилах. Спецификации разных деятельностей в составе потоков работ могут формулироваться в разных диалектах правил. Оркестровка потока работ выражается посредством продукционных правил (в диалекте RIF PRD). В продукционных правилах могут использоваться предикаты, определённые в других диалектах при спецификации деятельностей. Для спецификации управляющих конструкций потоков работ определяется пространство имён со специальными предикатами:

- *variable-definition* и *variable-value* для организации потоков данных на основе переменных и их значений;
- *parameter-definition* и *parameter-value* для организации входных и выходных параметров потоков работ и значений параметров;
- *end-of-task* – индикатор завершения работы деятельности.

Спецификации потоков работ в подавляющем большинстве случаев представляют собой композицию образцов управления [19] (последовательности, условий, разбиения, соединения и других), однако спецификации управляющих потоков могут быть и произвольными. Использование образцов позволяет упростить интеграцию за счёт типичных частей спецификации. Так, следующая спецификация определяет образец разбиения потока, в котором деятельности *B* и *C* выполняются одновременно после выполнения деятельности *A* [4] (используется презентационный синтаксис RIF [20]):

```
Group (
If Not(External(wkfl:end-of-task(A)))
Then Do (Act(A)
  Assert(External(wkfl:end-of-task(A))))
If And(Not(External(wkfl:end-of-task(B)))
  External(wkfl:end-of-task(A)))
Then Do (Act(B)
  Assert(External(wkfl:end-of-task(B))))
If And(Not(External(wkfl:end-of-task(C)))
  External(wkfl:end-of-task(A)))
Then Do (Act(C)
  Assert(External(wkfl:end-of-task(C))))
)
```

Реализации потоков работ могут либо разрабатываться при помощи трансляции спецификаций правил в языки конкретных систем, работающих с определёнными диалектами правил, либо выбираться из существующих релевантных потоков работ, их фрагментов, отдельных деятельностей и сервисов.

Заметим, что не все описанные выше принципы учитываются только использованием спецификаций на правилах. Для связывания со спецификациями дополнительной информации, такой, как описания Dublin Core или аннотационные свойства OWL [21], в языке предусмотрен механизм аннотирования.

Аннотации могут сопровождать любой класс конструкций RIF в спецификациях правил. Аннотации – это фреймы со свойствами, которые должны быть сохранены при любых манипуляциях спецификациями, но не добавляют семантики с точки зрения правил. Так, группа правил, приведённая выше, может быть аннотирована идентификатором *Split1* в текущем пространстве имён и фреймом, определяющим дополнительную информацию в слоте *type* со значением *ParallelSplit*:

```
(* Split1 Split1 [ rdf:type -> wf:ParallelSplit ] *)
( Group ...)
```

При реализации правил в конкретной системе вывода спецификации метаданных игнорируются. Тем не менее, они могут обладать семантикой, не зависимой от правил. Обычно аннотации в RIF определяются в терминах специализированного словаря, специфицирующего набор предопределённых

свойств. Состав метаданных в настоящем исследовании не ограничивается набором определённых свойств, а включает в себя более развитые описания. Необходимые метаданные в данном исследовании определяются в терминах онтологий, которые могут отличаться для разных предметных областей.

В следующем разделе описывается состав метаданных потоков работ, необходимый для организации семантического поиска потоков работ по метаданным. Спецификации связи данных и реализуемых методов с понятиями предметной области, требований к качеству требуют определения соответствующих онтологий.

3. Модель метаданных, ориентированная на семантический поиск реализаций потоков работ

В качестве словарей метаданных в спецификациях используются онтологии предметных областей, а также онтологии, определяющие свойства элементов потоков работ в различных ракурсах рассмотрения, таких как качество и происхождение данных и методов.

Аннотации, которые определяют метаданные, целесообразно связывать со следующими элементами потоков работ, выраженных правилами:

- потоки работ в целом;
- входные и выходные параметры потоков работ;
- деятельности внутри потоков работ;
- входные и выходные параметры деятельностей;
- переменные, определяющие потоки данных;
- отдельные правила или группы правил, определяющие фрагменты потока работ;
- группы правил, определяющие образцы потоков работ [19] (как это сделано на примере в разделе 2).

Для реализации семантических подходов к поиску потоков работ, релевантных решаемой задаче, в первую очередь, необходимо развивать

спецификации предметной области, в которой собирается коллекция методов. Поиск потоков работ и сервисов, отвечающих требованиям задачи, необходимо связывать с онтологией предметной области, которой принадлежит коллекция и в которой решается задача. Для этого элементы спецификаций, описывающие потоки работ, объявляются экземплярами классов понятий онтологии предметной области. Отнесение метаобъекта спецификации к классу понятия в терминах онтологий OWL реализуется посредством отношения *rdf:type* в RDF-графе. Для более сложных описаний в терминах онтологии метаобъекты могут становиться экземплярами неименованных классов RDF, определённых как подпонятия понятий онтологии, но без введения новых понятий и свойств в онтологию.

Связывание метаданных с потоками работ и поиск релевантных элементов потоков работ далее рассмотрим на примере. В работе [4] описывается задача составления портфелей ценных бумаг, временные ряды котировок которых слабо коррелируют друг с другом, и выбора лучшего из них по определённым критериям.

Для решения данной задачи разрабатываются спецификации потока работ, включающего следующие подзадачи:

- поиск максимальных портфелей-кандидатов со слабо зависимыми друг от друга временными рядами котировок бумаг;
- оценка бумаг, входящих в портфели, с точки зрения разных критериев, в частности, финансово-экономического;
- оценка портфелей по соответствующим критериям как обобщение оценок бумаг, входящих в них;
- обобщение нескольких критериев оценки портфелей в общую оценку и выбор лучшего портфеля.

Для реализации потока работ используются данные об истории цен на бумаги, принадлежность компаний индексу S&P 500 (индекс оценивается на основе данных о капитализации пятисот крупных американских компаний), оценка соотношения доходности и риска, мониторинг тональности

высказываний инвесторов об определённых бумагах. Оценка по разным критериям выполняется в потоке работ параллельными ветвями.

При разработке спецификации потока работ, решающего данную задачу, правила необходимо связывать со спецификациями предметной области. Предметная область данной задачи использует понятия фондовых рынков. Для описания предметной области используем онтологию на языке OWL:

```
Class(Portfolio)
ObjectProperty(includesSecurity)
  ObjectPropertyDomain(includesSecurity Portfolio)
  ObjectPropertyRange(includesSecurity Security)
ObjectProperty(hasMetric)
  ObjectPropertyDomain(hasMetric Portfolio)

Class(Security)
ObjectProperty(hasIdentifier)
  FunctionalObjectProperty(hasIdentifier)
  ObjectPropertyDomain(hasIdentifier Security)
  ObjectPropertyRange(hasIdentifier Ticker)
ObjectProperty(listedIn)
  ObjectPropertyDomain(listedIn Security)
  ObjectPropertyRange(listedIn StockMarketIndex)
ObjectProperty(hasRate)
  ObjectPropertyDomain(hasRate Security)
  ObjectPropertyRange(hasRate StockMarketRate)
ObjectProperty(hasMetric)
  ObjectPropertyDomain(hasMetric Security)
  ObjectPropertyRange(hasMetric Metric)
ObjectProperty(correlatesWith)
  ObjectPropertyDomain(correlatesWith Security)
  ObjectPropertyRange(correlatesWith Security)

Class(StockMarketRate)
ObjectProperty(onDate)
  FunctionalObjectProperty(onDate)
  ObjectPropertyDomain(onDate StockMarketRate)
  ObjectPropertyRange(onDate Date)

Class(Metric)
ObjectProperty(isMetricOfSecurity)
Class(Correlation)
  SubClassOf(Correlation Metric)
  SubClassOf(Correlation ObjectAllValuesFrom(isMetricOfSecurity Security))
Class(FinancialMetric)
  SubClassOf(FinancialMetric Metric)
Class(SocialMetric)
  SubClassOf(SocialMetric Metric)
```

Онтология[6] фондовых рынков, кроме прочего, определяет следующие необходимые в данной задаче понятия:

[6] http://ontology.ipi.ac.ru/ontologies/stockmarket.owl

- *Portfolio* – портфель, составленный из ценных бумаг определённого списка компаний, имеющий, с ним также могут быть связаны метрики оценки портфеля;
- *Security* – ценные бумаги компании, участвующие в фондовом рынке, у них есть идентификаторы, они могут принадлежать списку фондового индекса, оцениваются котировками, метриками надёжности, могут иметь зависимость от других бумаг;
- *StockMarketRate* – котировка бумаги, зависящая от времени;
- *Metric* – метрика для оценки надёжности ценной бумаги или портфеля; одной из метрик оценки надёжности бумаги является корреляция её котировки с другими бумагами.

Особо отметим, что представленная онтология определяет понятия и связи предметной области в отличие от спецификации концептуальной схемы [22], определяющей представление данных при решении задачи в потоке работ на правилах, хотя и онтология, и концептуальная схема используют выразительные средства языка OWL. Описания концептуальной схемы недостаточны для использования в метаданных о предметной области, так как многие понятия и отношения предметной области сведены в ней к примитивным типам данных (числовой, строковый). Фактически при этом понятия подразумеваются, но явно не описаны. Подробнее различия и связи онтологий и концептуальных схем предметных областей обсуждаются в [23].

Одновременно с онтологией предметной области для определения метаданных потоков работ используются другие онтологии, позволяющие выражать в метаданных различные аспекты описываемых элементов потоков работ.

Спецификации правил RIF чётко не разделяют определённый ими поток работ на элементы, принадлежащие его графовой структуре. Для спецификации одной деятельности или потока данных может использоваться набор правил, и по самим спецификациями не всегда возможно определить, какой графовой структуре принадлежит то или иное правило. Для связывания правил с видами

элементов потоков работ используется онтология структуры потоков работ[7], фрагмент которой приведён здесь:

```
Class(Workflow)
ObjectProperty(hasTask)
  ObjectPropertyDomain(hasTask Workflow)
  ObjectPropertyRange(hasTask Task)

Class(Task)
ObjectProperty(hasParameter)
  ObjectPropertyDomain(hasParameter Task)
  ObjectPropertyRange(hasParameter TaskParameter)
  InverseObjectProperties(isParameterOf hasParameter)
ObjectProperty(hasInputParameter)
  SubObjectPropertyOf(hasInputParameter hasParameter)
  ObjectPropertyDomain(hasInputParameter Task)
  ObjectPropertyRange(hasInputParameter InputParameter)
  InverseObjectProperties(isInputParameterOf hasInputParameter)
ObjectProperty(hasOutputParameter)
  SubObjectPropertyOf(hasOutputParameter hasParameter)
  ObjectPropertyDomain(hasOutputParameter Task)
  ObjectPropertyRange(hasOutputParameter OutputParameter)
  InverseObjectProperties(isOutputParameterOf hasOutputParameter)

Class(TaskParameter)
Class(InputParameter)
  SubClassOf(InputParameter TaskParameter)
Class(OutputParameter)
  SubClassOf(OutputParameter TaskParameter)

Class(ControlFlowPattern)
  SubClassOf(ControlFlowPattern Task)
Class(ExclusiveChoice)
  SubClassOf(ExclusiveChoice ControlFlowPattern)
Class(ParallelSplit)
  SubClassOf(ParallelSplit ControlFlowPattern)
  SubClassOf(ParallelSplit ObjectExactCardinality
    (1 hasInputParameter InputParameter))
  SubClassOf(ParallelSplit ObjectMinCardinality
    (2 hasOutputParameter OutputParameter))
Class(Synchronization)
  SubClassOf(Synchronization ControlFlowPattern)
  SubClassOf(Synchronization ObjectMinCardinality
    (2 hasInputParameter InputParameter))
  SubClassOf(Synchronization ObjectExactCardinality
    (1 hasOutputParameter OutputParameter))
Class(SimpleMerge)
  SubClassOf(SimpleMerge ControlFlowPattern)
```

В приведённом фрагменте онтологии структуры потока работ определены понятия:

- *Workflow* – потоки работ в целом, состоящие из наборов деятельностей;

[7] http://ontology.ipi.ac.ru/ontologies/wf.owl

- *Task* – деятельности, которые могут иметь входные и выходные параметры;
- *TaskParameter* – параметры деятельности;
- *ControlFlowPattern* – образцы управляющих конструкций потоков работ, и в частности, разбиение (*ParallelSplit*) и объединение (*Synchronization*) потоков.

Помимо этого, онтология определяет разновидности деятельностей, такие, как начало и завершение потока, понятие подпотока, другие образцы управляющих потоков и понятия структурных элементов потоков работ.

Для обеспечения надёжности данных и методов, наряду с модулями онтологии предметной области и структурных элементов потоков работ, спецификации метаданных необходимо пополнять также информацией в терминах специализированных онтологий, описывающих требования к происхождению данных и их качеству.

В качестве онтологии происхождения данных используется в соответствии с рекомендацией W3C онтология PROV-O [24]. В её основе лежат понятия агента (*Agent*), деятельности (*Activity*) и сущности (*Entity*). Агентами могут быть человек (*Person*), организация (*Organization*) или программа (*SoftwareAgent*). Вариации отношений их экземпляров друг с другом описывают различные события и ситуации, которые необходимо фиксировать при преобразовании, перемещении, изменении статуса данных. Например, метаданные об исходных данных, которые использовались процессом, выражаются отношением *used*, связывающим агента и деятельность; информация об инструменте, который был использован для генерации результата, выражается отношением *wasAttributedTo*, связывающим сущность и программу и так далее. Посредством такой онтологии можно задавать метаданные об авторстве и принадлежности данных и методов, прослеживать историю преобразования данных от первоначального источника до текущего состояния, сопровождать реальные данные и методы другой подобной информацией. Также в терминах онтологии происхождения выразимы

требования к средам воспроизведения, необходимым для реализации спецификаций.

Для определения требований к точности, полноте и другим аспектам качества данных, а также входных данных и ожидаемых результатов работы методов можно использовать онтологию качества данных DQ [25]. Она содержит набор факторов качества данных, определяемых измерениями в многомерном пространстве значений и метриками качества в этих измерениях. С одним объектом может одновременно быть связано несколько значений качества в разных измерениях. Примерами измерений качества являются требования полноты данных (Completeness), объёма данных (Data Volume), возраста данных (Timeliness), точности (Accuracy), целостности (Consistency), меры доверия (Confidence). Однако состав измерений и метрики, используемые для их реализации, могут сильно зависеть от предметной области исследования. Они согласуются и специфицируются сообществом, работающим в предметной области.

В терминах онтологий приведём метаданные, связанные со следующей спецификацией выходного параметра *ps* деятельности *getPortfolios* [4]:

```
( If Not(wkfl:end-of-task(getPortfolios))
  Then Do(Modify(external(wkfl:variable-value(ps svc:getPortfolios()))
    Assert(wkfl:end-of-task(getPortfolios))  )  )
)
```

Метаданные определяются перед соответствующим элементом спецификаций:

```
(* getPortfolios_ps
   getPortfolios_ps [ rdf:type -> wf:OutputParameter,
                      wf:isOutputParameterOf -> getPortfolios,
                      rdf:type -> pont:Portfolio,
                      prov:wasGeneratedBy -> getPortfolios,
                      prov:wasAttributedTo -> fsv:FinanceServices ]
*) ps
```

Данная спецификация метаданных определена для параметра *ps* в правиле, соответствующем деятельности потока работ. В первую очередь, она определяет в текущем пространстве имён уникальный идентификатор *getPortfolios_ps* данного элемента правила RIF. С этим идентификатором

связываются метаданные в терминах определённых выше онтологий (пространство имён *pont* соответствует онтологии предметной области, *wf* – онтологии структуры потоков работ, *prov* – онтологии происхождения данных, *fsv* – пространство спецификаций программы, используемой для генерации портфелей). Посредством RDF-отношения *type* определяется, что элемент с данным идентификатором является выходным параметром деятельности, связь с деятельностью определяется отношением *isOutputParameterOf* к метаобъекту с идентификатором *getPortfolios*. Он описывает объекты класса *Portfolio*, генерируемые программой *FinanceServices*, то есть возвращаемые деятельностью данные представляют собой портфели ценных бумаг, сгенерированные определённой программой. Идентификаторы *getPortfolios*, *FinanceServices* должны быть определены подобным образом в метаданных, связанных соответствующими спецификациями правил.

Таким образом, метаописание позволяет идентифицировать элементы спецификации в глобальном информационном пространстве, связывать спецификации правил с предметной областью, в которой решается задача, определять части правил, которые соответствуют элементам структуры потоков работ, а также семантически связывать элементы друг с другом с помощью выражений в терминах онтологий.

4. Поиск по метаданным

Спецификации правил, выражающие семантику поведения объектов в потоках работ, ортогональны сопровождающим их метаданным, поэтому правила и метаданные могут обрабатываться независимыми инструментами. Спецификации правил используются для реализации потоков работ в определённых системах в соответствии с семантикой используемых диалектов или для связывания со спецификациями существующих реализаций. Метаданные используются для поиска и предварительного связывания соответствующих элементов спецификаций.

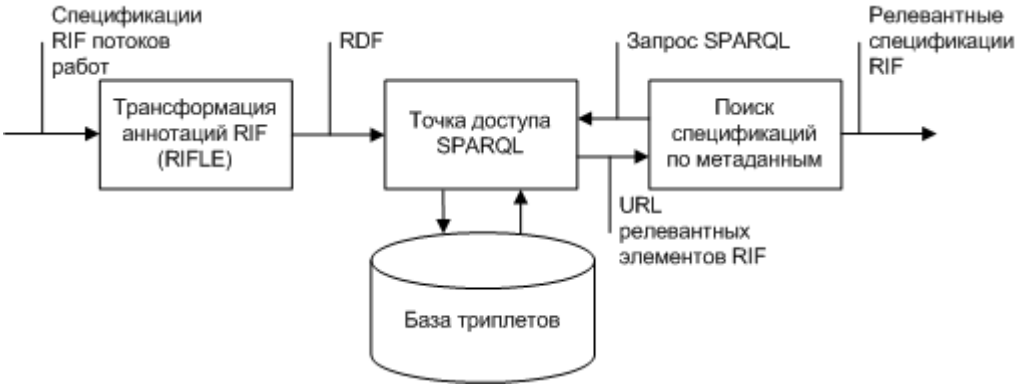

Рис 1: Архитектура подсистемы поиска спецификаций

На рис. 1 показана архитектура средств поиска потоков работ по метаданным. Спецификации RIF существующих потоков работ, аннотированные идентификаторами и метаданными в терминах онтологий, обрабатываются средствами библиотеки RIFle [27] для выделения в них аннотаций. Спецификации фреймов, содержащие значения метаданных, преобразуются в триплеты RDF в соответствии с рекомендациями W3C [26]. Фрейм с набором значений атрибутов $s[p_1->o_1\ p_2->o_2 \ldots\ p_n->o_n]$ соответствует триплетам, связанным с одним ресурсом $\{s\ p_1\ o_1;\ p_2\ o_2;\ \ldots\ p_n\ o_n\}$. Метаданные трансформируются в триплеты и через точку доступа SPARQL [28] сохраняются в базе триплетов с использованием библиотеки JENA [29]. База триплетов собирает в виде единого RDF-графа набор метаданных и идентификаторов, по которым можно установить, с какими именно элементами спецификаций потоков работ на правилах связаны определённые метаданные. В качестве RDF-словарей используется представленный выше набор онтологий, во многом зависящий от предметной области и определяющий состав метаданных. Поиск адекватных задаче спецификаций потоков работ и их фрагментов производится над базой триплетов, содержащей метаданные. Для поиска спецификаций в терминах онтологий, используемых в метаданных, формируются запросы SPARQL к данной точке доступа, имеющие целью найти идентификаторы спецификаций подходящих элементов потоков работ. По URI, соответствующим идентификаторам, выделяются спецификации RIF, релевантные задаче.

Следует заметить, что в аннотациях RIF нельзя использовать переменные, поэтому невозможно выразить неименованные узлы в RDF, и с URI, соответствующими идентификаторам в аннотациях RIF, связываются напрямую только значения свойств определённого типа данных или другие URI.

Для примера со спецификацией метода оценки ценной бумаги по финансово-экономическим критериям в базе триплетов связаны следующие метаданные:

```
res1:getSecurityFinancialMetrics
  rdf:type wf:Task;
  wf:hasInputParameter res1:finMetricPar;
  wf:hasOutputParameter res1:securityPar .
res1:securityPar
  rdf:type pont:Security;
  rdf:type wf:InputParameter;
  pont:hasMetric res1:finMetric .
res1:finMetric
  rdf:type pont:FinancialMetric;
  rdf:type wf:OutputParameter;
```

Они описывают деятельность *getSecurityFinancialMetrics* некоторого потока работ, имеющую входной параметр *securityPar*, являющийся описанием ценной бумаги, и выходной параметр *finMetric*, являющийся финансовой оценкой данной бумаги.

Запросы на языке SPARQL формулируются в соответствии с требованиями задачи, которая должна быть решена в предметной области, либо с требованиями спецификации потока работ, который необходимо реализовать посредством повторного использования существующих потоков работ, их фрагментов и доступных сервисов.

Требования к искомым потокам работ в коллекции научных методов могут затрагивать как семантику потока работ в целом или структуру и семантику элементов, выраженные в терминах онтологий, так и надёжность применяемых методов, достоверность используемых данных и получаемых результатов. Благодаря использованию разработанной модели метаданных в запросе могут быть учтены методы, процессы, применяющиеся в данной предметной области, семантика входных/выходных параметров потоков работ

и потоков на разных этапах решения задач, требуемая точность, контроль требований к реальным источникам данных и другие аспекты.

Решение научных задач предметной области может конструироваться из отдельных фрагментов потоков работ и из существующих сервисов. Фрагмент обработки данных может оказаться частью реализации потока работ, решающего в целом отличную задачу. Для этого требования в запросах формулируются как к потокам работ в целом, так и к параметрам деятельностей в составе потоков работ.

Для решения описанной выше задачи оценки портфелей ценных бумаг необходимо выбрать компоненты, реализующие оценки по каким-либо параметрам уже сформированных портфелей. При этом оцениваться должна каждая бумага в портфеле в отдельности, и применяется обобщающая оценка для всего портфеля. Зададим запрос SPARQL для поиска компонентов потоков работ, реализующих оценки по каким-либо параметрам выбранных ранее портфелей ценных бумаг.

```
select distinct ?task1 ?task2 where {
  ?in1 wf:isInputParameterOf ?task1 .
  ?in1 rdf:type pont:Security .
  ?in1 pont:hasMetric ?out1 .
  ?out1 wf:isOutputParameterOf ?task1 .
  ?out1 rdf:type pont:Metric .

  ?in2 pont:includesSecurity ?in1 .

  ?in2 wf:isInputParameterOf ?task2 .
  ?in2 rdf:type pont:Portfolio .
  ?in2 pont:hasMetric ?out2 .
  ?out2 wf:isOutputParameterOf ?task2 .
  ?out2 rdf:type pont:Metric .
}
```

По условию запроса необходимо найти связанные деятельности, одна из которых принимает на вход объекты ценных бумаг, вычисляет для них некоторую оценку, а вторая деятельность принимает на вход портфель, которому принадлежат данные бумаги, и вычисляет обобщающую оценку портфеля в целом. В результате обращения к базе триплетов система возвращает URI элементов потоков работ, удовлетворяющих представленному запросу:

```
<sparql xmlns=http://www.w3.org/2005/sparql-results#>
  <head>
  <variable name="task1"/>
  <variable name="task2"/>
  </head>
<results>
 <result>
  <binding name="task1"><uri>&res1#getPositiveTweetRatio</uri>
  </binding>
  <binding name="task2"><uri>&res1#computePortfolioTwitterMetrics</uri>
  </binding>
 </result>
 <result>
  <binding name="task1"><uri>&res2#getSecurityFinancialMetrics</uri>
  </binding>
  <binding name="task2"><uri>&res2#computePortfolioFinancialMetrics</uri>
  </binding>
 </result>
</results>
</sparql>
```

По возвращённым идентификаторам компонентов существующих потоков работ получаем доступ к их спецификациям и полным описаниям деятельностей, включая их метаданные:

- *getPositiveTweetRatio* – вычисляет тональность сообщений о ценной бумаге в Twitter;
- *computePortfolioTwitterMetrics* – на основе тональности сообщений о ценных бумагах вычисляет тональность отношения к содержащему их портфелю;
- *getSecurityFinancialMetrics* – вычисляет метрику надёжности ценной бумаги, учитывающую выгоду и риски на основе истории котировок;
- *computePortfolioFinancialMetrics* – для портфеля в целом, содержащего ценные бумаги, вычисляет обобщённую финансовую метрику.

Таким образом, найдены спецификации деятельностей, которые можно использовать повторно для реализации оценки ценных бумаг и портфелей при решении задачи выбора наилучшего портфеля. Для адаптации спецификаций потока работ к найденным компонентам необходимо использовать управляющую конструкцию разбиения потока для одновременного вычисления обеих оценок портфелей, а затем слияния потоков и вычисления обобщающей

оценки портфелей. Структура потока работ, использующего найденные спецификации, будет такой, как представлено на Рис. 2 [4].

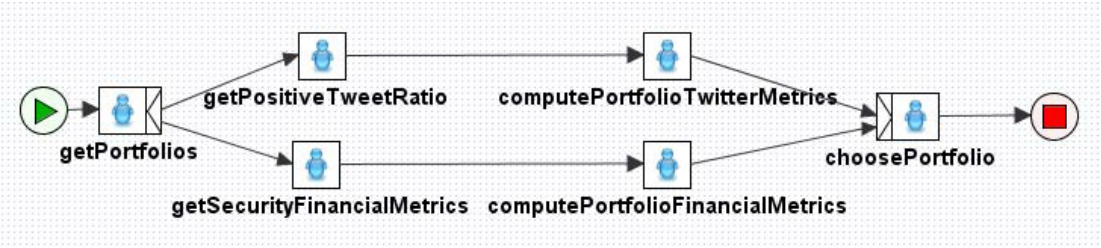

Рис 2: Поток работ для решения задачи выбора лучшего портфеля ценных бумаг

Помимо поиска, наличие метаданных о графовой структуре потоков работ, сопровождающих спецификации правил, позволяет применять эвристические методы оценки близости спецификаций потоков работ и методы проверки конформности спецификаций и их реализаций.

Представленную модель метаданных целесообразно использовать для проверки совместимости семантики данных и интероперабельности фрагментов при их объединении. Для этого необходимо проверять корректность включения подпроцессов по понятиям и сформулированным в метаданных требованиям к входным и выходным параметрам, соответствие семантики данных и входных/выходных параметров. Ограничение потока данных или постусловие выхода предыдущего компонента должно быть строже предусловия входа последующего компонента.

Заключение

В данном исследовании был разработан подход к сопровождению спецификаций потоков работ на правилах RIF метаданными в терминах онтологий OWL. Описана модель метаданных, включающая семантические описания компонентов потоков работ понятиями предметной области, требования к качеству данных и спецификации качества результатов работы деятельностей и потоков работы в целом, информацию о происхождении данных и методов в соответствии с рекомендациями W3C. Подход к поиску потоков работ и их фрагментов по метаданным проиллюстрирован на примере. При этом запросы могут включать требования ко всем аспектам, затрагиваемым

в разработанной модели метаданных. Использование языков правил для спецификации потоков работ даёт богатые возможности для повышения выразительности спецификаций и их повторного использования. В статье представлены требования, которые предъявляются к спецификациям потоков работ для их поиска и повторного использования и покрываются моделью потоков работ на правилах в совокупности с моделью метаданных.

Список литературы

1. *Н. А. Скворцов.* Подход к поиску потоков работ по метаданным // RCDL'2014. – Дубна: ОИЯИ, 2014.
2. RIF Overview. – W3C, 2013. – Адрес в интернете: http://www.w3.org/TR/rif-overview/
3. OWL 2 Web Ontology Language Document Overview (Second Edition) – W3C, 2011. – Адрес в интернете: http://www.w3.org/TR/owl-overview/
4. *L. Kalinichenko, S. Stupnikov, A. Vovchenko, D. Kovalev.* Multi-dialect Workflows // ADBIS'2014. – 2014. – LNCS 8716. – P. 352-365.
5. *Н. А. Скворцов, Д. О. Брюхов, Л. А. Калиниченко, Д. Ковалёв, С. А. Ступников.* Метаданные о научных методах для обеспечения их повторного использования и воспроизводимости результатов. // RCDL'2013. – Ярославль, 2014.
6. Aladin. Адрес в интернете: http://aladin.u-strasbg.fr/aladin.gml
7. Topcat. Адрес в интернете: http://www.star.bris.ac.uk/~mbt/topcat/
8. *N. A. Walton et al.* AstroGrid: A place for your science //Astronomy & Geophysics. – 2006. – Т. 47. – №. 3. – С. 22-24.
9. *D. Hull, K. Wolstencroft, R. Stevens, C.A. Goble, M.R. Pocock, P. Li, T. Oinn.* Taverna: A tool for building and running workflows of services // Nucleic Acids Research. – 34 (Web-Server-Issue). – 2006. – P. 729-732.
10. *C. A. Goble, D. C. De Roure.* myExperiment: social networking for workflow-using e-scientists // Proceedings of the 2nd workshop on Workflows in support of large-scale science. – ACM, 2007. – С. 1-2.

11. Wf4Ever project. – Адрес в интернете: http://www.wf4ever-project.org/
12. *C. Wroe, C. Goble, A. Goderis, P. Lord, S. Miles, J. Papay ; P. Alper, L. Moreau.* Recycling workflows and services through discovery and reuse // Concurrency and Computation: Practice and Experience. – 2007. –Vol. 19, No. 2. – C. 181-194.
13. *C. Tejo-Alonso et al.* Metadata for web ontologies and rules: Current practices and perspectives // Metadata and Semantic Research. – Springer Berlin Heidelberg, 2011. – C. 56-67.
14. *H. D. Burkhard, M. M. Richter.* On the notion of similarity in case based reasoning and fuzzy theory. // Soft computing in case based reasoning. – 2000.
15. *R. Bergmann, A. Fressmann, K. Maximini, R. Maximini, T. Sauer.* Case-based support for collaborative business. // Advances in CBR. – Vol. 4106. – Springer, 2006. - P. 519-533.
16. *D. B. Leake, J. Kendall-Morwick.* Towards Case-Based support for e-Science workflow generation by mining provenance. // Advances in CBR. – 2008. – P. 269-283.
17. *R. Bergmann, Y. Gil.* Retrieval of semantic workflows with knowledge intensive similarity measures // Case-Based Reasoning Research and Development. – Springer Berlin Heidelberg, 2011. – C. 17-31.
18. *W.M.P.Van der Aalst.* Process mining: Discovery, Conformance and Enhancement of Business Processes. – Heidelberg: Springer, 2011.
19. *N. Russell, A.H.M. ter Hofstede, W.M.P. van der Aalst, and N. Mulyar.* Workflow Control-Flow Patterns: A Revised View. – BPM Center Report BPM-06-22, BPMcenter.org. – 2006.
20. RIF Production Rule Dialect (Second Edition). – W3C, 2013. – Адрес в интернете: http://www.w3.org/TR/rif-prd/
21. OWL Web Ontology Language Reference. – W3C, 2004. – Адрес в интернете: http://www.w3.org/TR/owl-ref/

22. *L. A. Kalinichenko, S. A. Stupnikov, A. E. Vovchenko, D. A. Kovalev.* Conceptual declarative problem specification and solving in data intensive domains // Informatics and Applications. – Vol. 7, Iss. 4. – Moscow: IPI RAS, 2013.
23. *А. Е. Вовченко и др.* От спецификаций требований к концептуальной схеме // RCDL'2010. – Казань: КФУ, 2010. – С. 375-381.
24. The PROV Ontology. W3C Recommendation. – W3C, 2013. – Адрес в интернете: http://www.w3.org/TR/prov-o/
25. *S. Geisler, S. Weber, C. Quix.* Ontology-based data quality framework for data stream applications. // Proc. of the 16th International Conference on Information Quality (ICIQ-11). – 2011.
26. RIF RDF and OWL Compatibility. – W3C, 2013. – Адрес в интернете: http://www.w3.org/TR/rif-rdf-owl/
27. RIFle. – Адрес в интернете: https://bitbucket.org/fundacionctic/rifle/wiki/Home
28. SPARQL Query Language for RDF. – W3C, 2008. – Адрес в интернете: http://www.w3.org/TR/rdf-sparql-query/
29. Apache JENA. – Адрес в интернете: http://jena.apache.org/

The Metadata Model for Semantic Search for Rule-Based Workflow Implementations

N. A. Skvortsov
A. E. Vovchenko
L. A. Kalinichenko
D. A. Kovalev
S. A. Stupnikov