Интеграция и поиск информации в гетерогенных динамических информационных массивах с помощью онтологий

© А.В. Жучков[а], С.А. Арнаутов[а], Н.В. Твердохлебов[а], С.В. Голицын[б], И.Г. Стриж[в]

(а) Институт химической физики им. Н.Н. Семенова РАН, 119991 Москва, ул. Косыгина, 4

(б) Автономная некоммерческая организация «Телекоммуникационный центр «Наука и общество», 119991 Москва, ул. Косыгина, 4

(в) Московский государственный университет

arnautov@chph.ras.ru

Аннотация

Одним из наиболее перспективных направлений решения проблемы интеграции разнородной информации являются технологии семантического связывания на основе онтологий. В настоящей работе обсуждаются некоторые аспекты имплементации этой технологии в рамках существующего набора разнообразных коллекций в составе электронной библиотеки.

1 Введение

Одним из актуальных и приоритетных направлений в области естествознания, на сегодняшний день, являются биомедицинские и иммунологические исследования, направленные, в частности, на разработку новых иммунологических препаратов. Благодаря интенсивному развитию и применению молекулярно-биологических методов, в последние годы в этой области наблюдается резкое увеличение объема информации. Получаемые данные, представляющие собой, главным образом, последовательности нуклеиновых кислот, соответствующих гену, его регуляторной области, мотиву или сайту связывания, накапливаются в более 500 различных национальных и международных базах данных, объем которых достигает сотен Гигабайт (SWISS-PROT, PIR-PSD, Medline) и продолжает увеличиваться экспоненциально. Необходимая информация сосредоточена также в многочисленных библиографических базах данных; в локальных базах, содержащих результаты клинических наблюдений и результаты биологических исследований *in vitro*. В связи с тем, что многочисленные источники информации, необходимой для эффективной работы, в частности, иммунологов, создаются в разных местах специалистами различного профиля с использованием различных технологий и инструментария, речь идет о гетерогенных слабоструктурированных динамически изменяющихся информационных массивах большого объема. Мы столкнулись с этой проблемой в ходе выполнения межведомственной научно-технической программы «Вакцины нового поколения и диагностические системы будущего», направленной на разработку новых иммунологических препаратов. Число участников программы превышает 90 организаций, а объем данных - несколько гигабайт. Информационная система, созданная для информационного обеспечения этой программы функционирует на базе Южной Московской Опорной Сети (ЮМОС) и включает корпоративную сеть и электронную библиотеку (ЭБ) [1]. В настоящее время, в рамках проводимой работы, актуальной задачей представляется систематизация существующих динамических массивов, а также создание инструментария, позволяющего пользователю эффективно ориентироваться в информационном пространстве и осуществлять поиск необходимой информации.

В результате анализа литературы был сделан вывод о том, что наиболее перспективным направлением решения проблемы интеграции разнородной информации являются технологии семантического связывания на основе онтологий [2].

2 Аспекты имплементации

2.1 Определение онтологии и требования к инструментарию

Существует несколько определений термина «онтология». В данной работе под онтологией понимается эксплицитная (явная, вербализованная, формальная) спецификация некоторой концепции [3]. При этом речь идет об онтологиях уровня предметной области (domain specific) или конкретной задачи (task specific).

Для эффективного использования онтологий необходимо наличие инструментов, позволяющих:

- создавать и модифицировать онтологии;
- импортировать онтологии из других источников;
- объединять онтологии различного происхождения;
- автоматически создавать онтологии для источника данных с использованием метаданных этого источника;
- осуществлять различные другие операции над онтологиями;
- поддерживать аннотирование и версионность онтологии;
- связывать концепты онтологии с различными типами данных;
- обеспечивать многоязычную поддержку.

Очевидно, что в масштабах одного, даже достаточно крупного проекта, нет возможности имплементации всех технологий с использованием семантики и онтологий. В настоящей работе предпринята попытка сформулировать предлагаемый нами подход, и изложены первые результаты..

2.2 Онтологии в гетерогенной ЭБ

Поскольку описываемая ЭБ создавалась уже несколько лет, она содержит разнородные коллекции, содержащие данные и информацию из следующих областей:

- данные молекулярно-биологического и генетического характера: аминокислотные последовательности белков и нуклеотидные последовательности генов и регуляторных мотивов, участвующих в иммунном ответе;
- данные о природных и химически синтезированным соединениях, оказывающих влияние на иммунный ответ организма (иммуномодуляторы и адъюванты);
- данные об инфекционных агентах и аллергенах: их таксономия, происхождение, распространение в природе и т.п.;
- данные о вакцинах: природе их иммуногена, компонентах, способах производства и применения, рекомендации к вакцинации и критерии оценки эффективности;
- данные по диагностическим тест-системам, основанные на регистрации иммунного ответа организма в ответ на какой-либо антиген;
- библиографические данные;
- вспомогательные данные (методические указания, объекты исследования и др.).

На первом этапе семантико-онтологический подход предлагается использовать для предоставления пользователю возможности а) эффективно ориентироваться в содержании той или иной коллекций ЭБ; б) оптимизировать поиск необходимых данных; в) создавать собственные онтологии при формировании т.н. «авторских наборов данных»; г) генерировать новое знание в рамках данной понятийной сети.

2.3 Онтология как инструмент ориентации пользователя ЭБ

Онтологии позволяют не только и не столько структурировать содержимое информационных массивов (они уже структурированы, эту функцию выполняют модели баз данных и/или метаданные), сколько получить полное представление о содержимом ЭБ. Примечательно, что поскольку любая онтология, как правило, является авторским набором концептов и реляций, то разные эксперты-пользователи в зависимости от интересующей их проблемы могут создавать различные онтологии для одного и того же авторского набора. Иными словами, вместо одного общего «содержания», по которому пользователь пытается найти необходимую ему информацию, в результате создания онтологий мы имеем дело с несколькими детализированными «подразделами». Таким образом, онтологии могут позволить пользователю по новому взглянуть на имеющиеся данные и рассматривать их, возможно, даже в ином контексте: применяя ту или иную онтологию к ЭБ, мы можем получить разные знания. Например, если брать онтологию, основанную на классификации природы вакцин, то можно найти каким образом вакцины могут быть получены. А если брать онтологию по заболеваниям, то – какие вакцины могут быть использованы для лечения того или иного заболевания.

2.4 Совместная работа над онтологией как процесс генерации нового знания

На этот процесс можно взглянуть с другой стороны. Поскольку специализированные онтологий по определенным тематикам создаются исходя из уже существующих коллекции, в совокупности они образуют библиотеку слабосвязанных частных онтологий. Их можно объединить в сводную онтологию, но это достаточно искусственная операция. Но если рассматривать деятельность участников программы как функционирование понятийной сети, имеющей целью достижение результатов данной программы, то результатом их работы должно стать некое общее

понимание, видение той предметной области, в рамках которой эта сеть функционирует. Нам представляется, что конкретной формой такой совместной деятельности может быть создание и развитие единой онтологии данной ИС (как отражения данной предметной области). Речь идет о генерации нового знания, формализуемого в онтологию, разделяемую всеми участниками данной ПС. Естественно, ПО, создаваемое для реализации такой функциональности, должно обладать соответствующими возможностями.

2.5 О размерах онтологии

Очевидно, что разработка онтологического описания определенной предметной области, невозможна без выработки, формирования набора или словаря терминов (тезауруса). Таким образом, важным моментом в создании онтологии является использование концептов, имеющих четкое определение. Более того, можно рекомендовать использовать в качестве концептов общепринятые, общеупотребительные, базовые понятия и термины, а также вовлекать общепринятые схемы и классификации в создаваемые онтологии. Большой исторический опыт по созданию классификационных систем в области биологии и медицины свидетельствует о том, что как только создаваемые классификации становятся большими, детализированными и разветвленными, они теряют гибкость и их уже трудно поддерживать. Для нужд данной ИС онтологии намеренно разрабатывались предельно компактными и обозримыми, чтобы сделать работу с ними удобной для пользователей. Например, анализ определенного концепта и связанных с ним данных позволяет формировать новые, дополнительные концепты онтологий и устанавливать новые связи-взаимоотношения.

2.6 G-Ontology – оригинальный инструмент для работы с онтологиями

В качестве инструмента для работы с онтологиями было разработано оригинальное программное обеспечение (ПО) G-Ontology. Созданное ПО совместимо с созданным ранее ПО Gozelle, которое обеспечивает функциональность ИС в целом (поиск данных, создание и модификация данных; различные интерфейсы взаимодействия со сторонними источниками данных; работа с авторскими и распределенными наборами данных; работа с метаданными; различные фильтры и конверторы для работы с гетерогенными распределенными электронными библиотеками; различные; разнообразные поисковые средства) [4]. G-Ontology сочетает в себе графический модуль, что позволяет эксперту и пользователю создавать собственные онтологии в удобном для него виде, а также специальный формат для описания концептов и реляций онтологий. G-Ontology поддерживает групповую работу с онтологией, в частности, хранит записи о соавторах (в формате VCard), а также вносимые изменения. История создания онтологии позволяет проанализировать весь процесс работы над ней и, при необходимости, вносить требуемые коррективы.

В состав онтологии также входит каталог ресурсов, которые она описывает. Для обеспечения возможности создания иерархической онтологии и навигации по ней, наряду с каталогом ресурсов, в G-Ontology присутствует каталог онтологий.

Текущая версия G-Ontology позволяет пользователю не только строить собственные онтологии, но и совершать определенные операции над ними, в т.ч.:

- представление концептов и реляций на различных языках (русском, английском, немецком и пр.);
- объединение двух и более онтологий по общему концепту;
- возможность создания многоуровневой онтологии;
- аннотирование концептов онтологии на нескольких языках;
- фильтрация сегментов онтологии;
- возможность просмотра и модифицирования метаданных концепта;
- возможность просмотра и модифицирования данных связанных с определенным концептом;
- автоматическая фильтрация видимой части онтологии на основе различных критериев (по определенным авторам, дате создания концепта или реляции);
- авторская фильтрация, осуществляемая пользователю вручную;
- создание, модификация, удаление и аннотирование типов реляций, определенных для данной онтологии;
- достаточно гибкий синтаксис для «привязывания» данных к концепту онтологии.

Разрабатываемая в настоящее время новая версия ПО будет поддерживать автоматическое создание онтологии для источника данных, используя семантические метаданные ресурса, частотные словари и тезаурусы предметных областей.

3. Заключение

Использование онтологий открывает новые возможности при создании современных электронных библиотек. Особенно ярко преимущества этой технологии проявляются в больших распределенных гетерогенных информационных системах. Одним из важнейших следствий является возможность генерации нового знания, что означает эволюцию ЭБ в направлении баз (библиотек) знаний.

Литература

[1] Жучков А.В., Твердохлебов Н.В., Арнаутов С.А., Голицын С. От информационной системы проекта (учреждения) к электронной библиотеке в понятийной сети. Труды V Всероссийской объединенной конференции «Технологии информационного общества – интернет и современное общество». Санкт-Петербург, 25-29 ноября 2002 г., с.91-94

[2] Stevens R., Goble C.A., Bechhofer S. Ontology-based Knowledge Representation for Bioinformatics. Brief. Bioinformatics, 2000, 1(4):398-416

[3] T.R. Gruber. A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2), 1993.

[4] Жучков А.В., Арнаутов С.А, Твердохлебов Н.В., Голицын С.В., Стриж И.Г. Новые технологии для понятийных сетей, создаваемых в рамках МНТП «Вакцины нового поколения и диагностические системы будущего» «Электронные библиотеки», 2003, т.6, вып. 6 (http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2003/part6/ZATGS)

Ontology-driven data integration and data mining in dynamic heterogeneous data sources

© Joutchkov A.[1], Arnautov S.[1], Tverdokhlebov N.[1], Golisyn S.[2], Strizh I.[3]

[1] Semenov Institute of Chemical Physics of RAS, 119991 Moscow, Kosygina 4

[2] Telecommunication Centre “Science and Society”, 119991 Moscow, Kosygina 4

[3] Moscow State Unniversity, Department of Biology

arnautov@chph.ras.ru

Ontology-driven semantic coupling seems to be the most promised technology to integrate heterogeneous data sources.

We discuss here about some problems of this technology implementation to integrate an existing set of different collections in framework of an electronic library.