УДК 004.822

С.П. Некрашевич, Д.В. Божко
Донецкий государственный институт искусственного интеллекта, Украина

Представление данных в Интернет на основе семантических сетей

В данной статье рассматриваются современные модели представления и обработки данных в Интернет. Проанализирована реляционная модель представления данных и некоторые модели представления знаний. Предложены средства и методы повышения семантической связности информации, на основе которых происходит поиск в сети Интернет. Приведено описание алгоритма перехода от реляционной модели данных к семантической сети. Предложен язык формализации семантической сети. Описаны возможные способы применения разработанной модели.

В настоящее время основным источником структурированных данных в Интернет являются базы данных, построенные на основе реляционной модели данных. Для них характерны следующие свойства:

– простота моделирования концептуального представления предметной области;

– четко определенная модель данных и алгебра отношений между понятиями (концептами);

– специализированный язык запросов SQL для обработки данных (поиск, модификация, выполнение транзакций и пр.);

– использование в качестве стандарта при построении систем «клиент-сервер» для коммерческого использования и др.

Однако при использовании этой модели для создания интеллектуальных систем возникают определенные проблемы [1], например:

– отсутствие адекватного способа описания смысла (семантики) данных, представляющих понятия предметной области;

– отсутствие средств представления знаний и поиска (вывода) новой информации на их основе;

– ограниченность типов данных и отношений между реляционными таблицами, жесткая зависимость запросов от логической модели данных;

– высокая стоимость сопровождения существующих и унаследованных баз данных, необходимость частых модификаций для реализации новой функциональности;

– слабая масштабируемость базы данных при выполнении нетривиальных запросов поиска информации и выполнения задач исследования данных.

Цель работы

Для устранения вышеуказанных проблем и повышения уровня абстракции и интеллектуальности реляционной модели предлагается использовать семантическое описание концептов и отношений между ними на основе моделей представления знаний в качестве вспомогательной информации к существующим запросам и схемам представления данных. Запрос к базе данных осуществляется на основе

стандартного языка SQL и отдельной информации, которую пользователь сервиса указывает дополнительно для уточнения запроса. Таким образом, становится возможным параллельное сопровождение баз данных с предоставлением новых высокоуровневых интеллектуальных сервисов.

Выбор модели представления знаний

В результате анализа существующих моделей представления знаний [2] можно выявить наиболее подходящую модель для последующей реализации интеллектуальных сервисов на основе реляционных баз данных.

Продукционная модель, основанная на правилах, позволяет представить знания в виде предложений типа «если <условие>, то <действие>». Её применение может быть оправдано только для описания взаимодействия концептов предметной области, представление же семантики данных должно быть четко определено в форме, которая может оказаться сложной для восприятия пользователем сервиса.

Фреймовая модель, получившая дальнейшее развитие в объектной и компонентной моделях [3], требует определенной квалификации пользователя в формализации представления фреймов (описание структур, ролей, сценариев и ситуаций предметной области).

Семантическая сеть, в основе которой находится математическое понятие графа, вершины которого – понятия предметной области, а дуги – отношения между ними, наиболее удобна для моделирования пользователем запросов, а также уточняющей информации к существующим запросам SQL. Проблема поиска решения в базе знаний типа семантической сети сводится к задаче поиска фрагмента сети, соответствующего некоторой подсети, отражающей поставленный запрос к базе.

Формализация данных и знаний на основе семантических сетей

Дальнейшим развитием технологий Интернета и World Wide Web является Semantic Web. Этой технологии присущи следующие свойства:

- аннотирование данных, накопленных в Интернете за время его развития;
- мета-описание данных на основе онтологий;
- взаимное отображение онтологий;
- интеллектуальные сервисы, учитывающие и понимающие семантику данных.

Наибольшее распространение в Semantic Web получили следующие форматы описания данных:

- ODL – стандарт ODMG объектно-ориентированных БД;
- RDFS (Resource Definition Framework Schema) – стандарт позволяет описывать схемы классов и их свойств с учетом отношений между ними;
- OWL (Web Ontology Language) – специализация RDFS, ориентированная на описание предметных онтологий.

Целесообразно для формального представления данных и знаний семантической сети использовать, соответственно, формат RDFS – для представления данных и OWL – для представления знаний.

Использование в реляционной модели семантически аннотированных данных позволяет:

- представлять информацию в унифицированном формате;
- обеспечить синтаксическую интероперабельность сервисов на основе различных схем формата XML;
- обеспечить семантическую интероперабельность на основе онтологий.

Все вместе это позволит повысить уровень абстракции модели предметной области и повысить интеллектуальность сервисов, предоставляемых пользователю сети Интернет.

В вопросе интеграции (точнее, технической интероперабельности) распределенных репозиториев данных все большую силу набирает технология Web-сервисов как средства предоставления унифицированного, платформенно-независимого интерфейса для удаленного доступа к информационным ресурсам. В данном контексте Web-сервис выступает в роли автономного приложения, которое предоставляет средства доступа к информации внешним клиентам через набор предоставляемых им услуг. Технология Web-сервисов базируется на таких открытых XML-стандартах, как:

- SOAP (Simple Object Access Protocol) – XML-протокол для удаленного вызова методов Web-сервисов;
- UDDI(Universal Description, Discovery and Integration) – описывает модель данных, предназначенную для каталогизации и обнаружения услуг, предоставляемых Web-сервисами;
- WSDL (Web Services Description Language) – язык описания интерфейсов Web-сервисов.

Формирующиеся дополнения к ним, например, WSCoordination/WS-Transaction (транзакции), WSSecurity (безопасность), WS-Routing (маршрутизации сообщений) и т.д., призваны расширить возможности этой платформы в удовлетворении требований задач интеграции приложений. В рамках инициативы WS-I разрабатываются примеры прикладных решений, предложения и дополнительные требования, призванные гарантировать совместимость решений разных поставщиков.

Сервисы на основе семантически аннотированных данных

Во многих случаях интеграция информационных ресурсов требует комбинирования обращений более, чем к одному Web-сервису для реализации пользовательского запроса [4]. Таким образом, Web-сервисы должны иметь возможность поддерживать взаимодействие с другими приложениями в дополнение к стандартным процедурам обработки данных. Более того, процесс предоставления агрегированной распределенной информации может включать в себя разбиение на набор взаимосвязанных этапов обработки данных, взаимодействие ряда Web-сервисов, вмешательство людей в процесс обработки пользовательских запросов и другие элементы прикладной логики.

Поэтому процесс сбора и интеграции структурированных данных может представлять собой логически сложную композицию обращений к хранилищам

информационных сущностей посредством интерфейсов Web-сервисов – определять автоматизированный поток обработки данных.

Для описания композиций Web-сервисов на данный момент различными ассоциациями предлагается ряд стандартов. Среди них можно отметить следующие языки описания автоматизированных потоков работ, участниками которых являются Web-сервисы:

– WSFL (Web Services Flow Language) – позволяет определять композиции Web-сервисов в виде графовой модели рабочего процесса;
– BPML (Business Process Modeling Language) – определяет блочную модель композиции Web-сервисов;
– BPEL4WS (Business Process Execution Language For Web-Services) – представляет собой гибрид блочной и графовой моделей описания взаимодействий Web-сервисов.

Эти языки позволяют описывать композиции Web-сервисов, что позволяет определять сложные, распределенные процессы по извлечению, обработке и интеграции информации.

Итак, мы можем выделить метод осуществления процесса сбора и интеграции распределенных данных, который базируется на трех технологиях:

– объектные репозитории данных, соответствующие некоторым предметным областям;
– механизм Web-сервисов как средство построения внешних интерфейсов к таким репозиториям;
– аппарат рабочих процессов как средство управления обработкой и интеграции информационных потоков.

Реляционная модель данных как семантическая сеть

Понятие «интеграция распределенных данных» подразумевает, как правило, интеграцию информационных ресурсов, которые расположены в уже существующих распределенных репозиториях [5]. В настоящее время большая часть информационных хранилищ представлена реляционными базами данных. Поэтому первая задача, возникающая на пути решения проблемы семантически обоснованной интеграции информационных ресурсов – это представление данных, описанных реляционной моделью, семантически более богатым способом [3].

Таким образом, необходимо наличие механизмов, позволяющих выделить из реляционной модели данных объектную модель и реализовать адаптер для работы с данными существующего хранилища информационных ресурсов через объектные интерфейсы доступа.

Была использована методика, которая опиралась на реинжениринг реляционных схем существующих реляционных хранилищ данных, создание соответствующих объектных схем данных и возможности программного комплекса, базирующегося на Java-технологиях, которые позволяют сформировать «объектную» надстройку над имеющимся реляционным хранилищем информационных ресурсов для того, чтобы работать с его данными посредством технологий Semantic Web.

Если определить реляционную модель в области понятий онтологии, то получим онтологию с жестко ограниченным количеством типов отношений между

понятиями предметной области. Этот факт мешает получить описание предметной области с использованием более обширного числа типов связей. С применением семантической модели получаем возможность использовать большее число типов связей.

Весь процесс перехода от реляционной модели к семантической схематично показан на рис. 1.

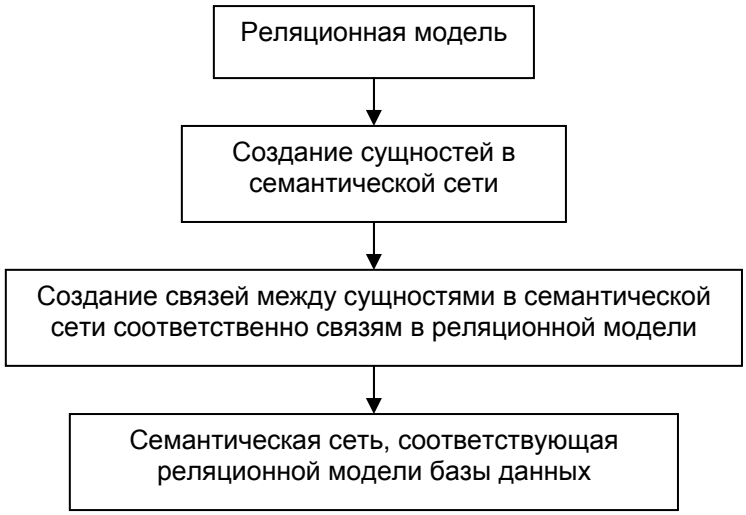

Рисунок 1 – Общая схема перехода от реляционной модели к семантической

На первом этапе, при переходе от реляционной модели к семантической сети, нужно определить предметную область, в которой будет работать создаваемая сеть. Так как реляционная модель имеет свою реализацию в виде БД с её выделенными в таблицы понятиями и установленными отношениями, то в семантическую сеть добавляются эти понятия. Далее «копируются» отношения. Типы реляционных отношений в базе данных можно привести к отношению типа «это». Такое приведение становится возможным, если понятия базы данных представить как сложные, т.е. состоящие из более простых терминов. Под более простыми терминами здесь предлагается понимать поля таблиц базы. Таким образом, получаем онтологию, которая содержит понятие и его определение.

Использование семантической сети, которая отражает один в один реляционную модель, не принесет никакой функциональной выгоды, а только приведет к потере времени на её создание и обработку. Таким образом, следующим шагом к повышению описательной способности сети должно быть её расширение.

Для расширения полученной семантической сети предлагается семантическая сеть, созданная на основе реляционной модели базы и расширенная при помощи эксперта по интересуемой предметной области.

Процесс расширения сети может быть автоматизирован, так как для этого не требуются инженерные навыки, а только необходимые знания в предметной области.

Эффективность использования предлагаемой модели поиска и представления данных заключается в следующем:

– повышении релевантности ответов за счет использования более обширного описания предметной области;
– получении ответов на запрос пользователя с учетом связанных с указаными в запросе понятиями и ограничениями;
– интерпретации ограничений и понятий в приедметной области;
– повышении скорости обработки сложных запросов как следствие вторичного использования выделеных связей в расширенной семантической сети;
– уменьшении трафика передачи данных между сервером и клиентом при выполнении сложных запросов с использованием больших объемов данных;
– предоставлении унифицированного интерфейса для эффективного доступа к информационным системам.

Модель представления и обработки данных на основе семантических сетей

Основные этапы формирования объектного репозитория схематически представлены на рис. 2. Для выделения объектной схемы реляционных баз данных внешних систем в рамках разработанной методики необходимо выполнить определенную последовательность действий.

1. Формирование ER-схемы для БД целевой системы. На первом этапе необходимо получить схему существующей реляционной базы данных для того, чтобы впоследствии преобразовать ее к объектной схеме, внеся дополнительное семантическое наполнение и структуризацию. Выделение ER-схемы существующей БД целевой системы можно выполнить следующими программными средствами:
– MS Visio 2000/2002/2003 (позволяет построить системную ER-схему БД в ER-нотации);
– IBM RROSE 2000/2002/2003 (позволяет с помощью модуля Data Modeller сформировать системную ER-схему целевой БД).

2. Формирование UML-диаграммы классов по ER-схеме целевой системы. Второй этап в построении объектного репозитория над реляционной базой данных – это преобразование полученной ER-схемы данных к первому приближения OWL-модели информационных ресурсов Semantic Web. В качестве этого первого приближения удобно использовать UML-диаграмму классов. Формирование UML-диаграмм классов по ER-схемам можно выполнить следующими программными средствами:
– MS Visio 2000/2002/2003 (не умеет преобразовывать ER-схемы в UML-диаграммы классов, ввиду чего требуемое преобразование необходимо выполнить «руками», имея в редакторе две эти схемы);
– Poseidon for UML (не умеет преобразовывать ER-схемы в UML-диаграммы классов, ввиду чего требуемое преобразование необходимо выполнить «руками», имея в редакторе две эти схемы);
– IBM RROSE 2000/2002/2003 (представляет ER-схему в UML-нотации по собственной методике).

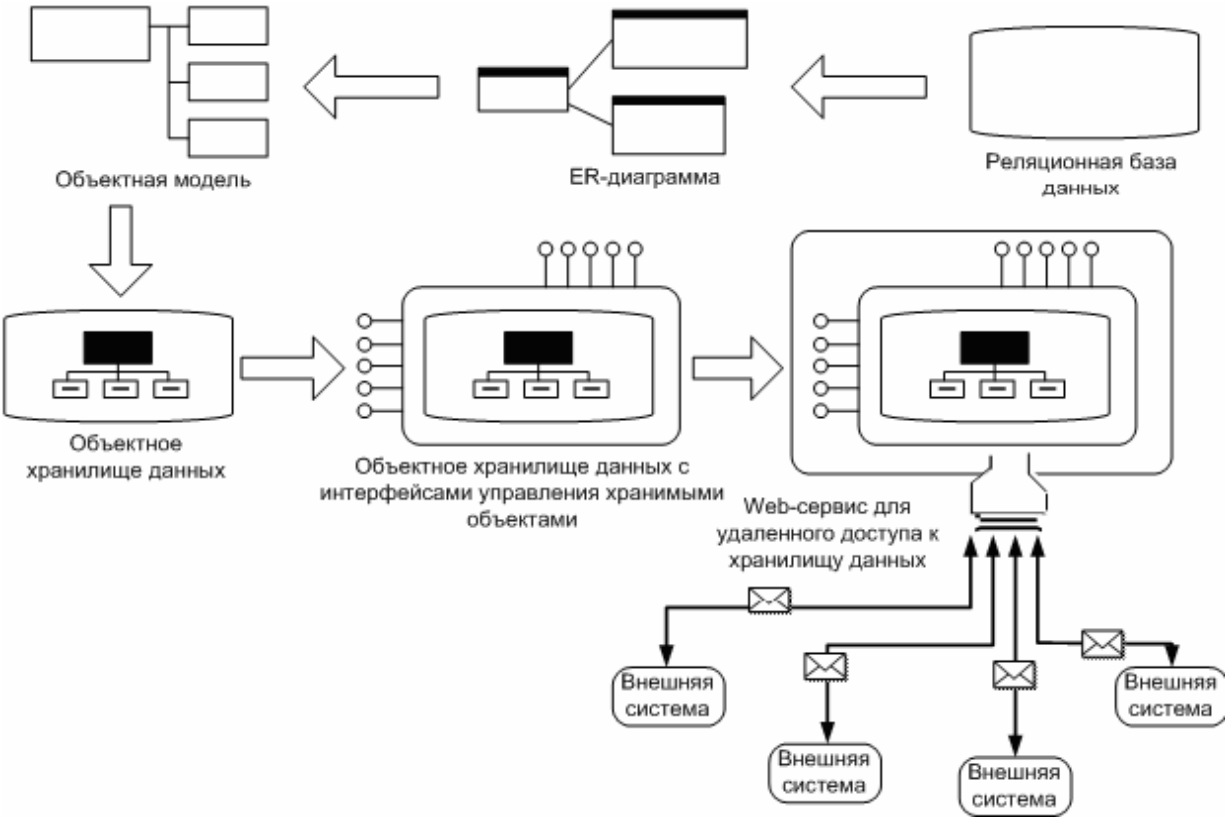

Рисунок 2 – Модель представления и обработки данных
на основе семантических сетей

3. На следующем этапе нам необходимо представить полученную UML-диаграмму классов в некоторой промежуточной, схемо-независимой форме для последующего преобразования к модели данных OWL.

Общая схема обработки реляционных запросов с использованием семантической сети показана на рис. 3.

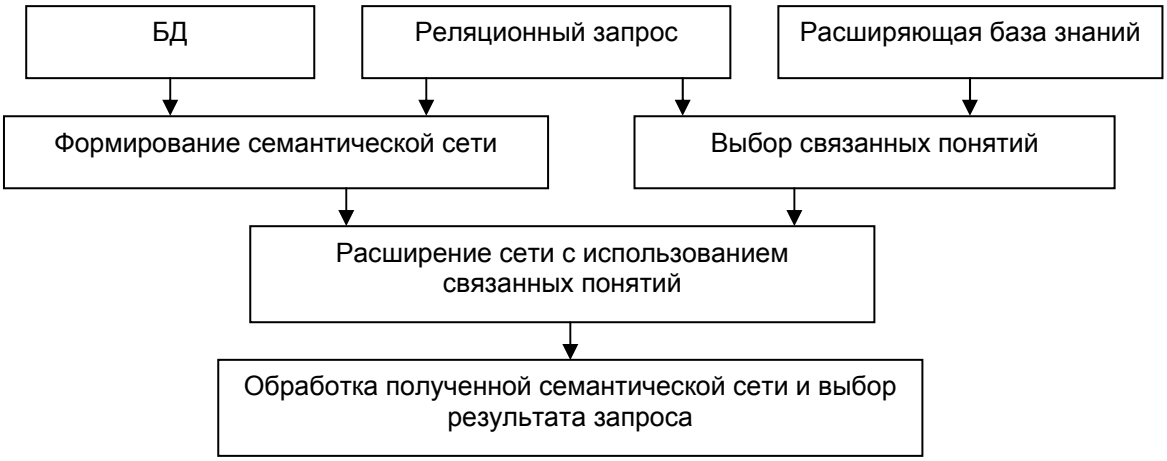

Рисунок 3 – Схема обработки запросов к реляционной базе с использованием
семантической сети

4. Преобразование UML-диаграмм классов в промежуточной форме в OWL-структуры. На данном этапе подготовлены все необходимые входные артефакты для построения первого варианта OWL-структуры, описывающей схему данных объектной надстройки над реляционной базой данных.

5. Формирование прикладной OWL-структуры. После выделения первого приближения OWL-модели данных объектного репозитория необходимо выполнить доработку полученного первого варианта схемы до семантически более корректной формы. На данном этапе предполагается:

– доработка OWL-структуры: введение дополнительной иерархии классов и их свойств;

– введение системных классов технологической платформы, необходимых адаптеру объектного репозитория;

– формирование OWL-схемы, согласованной с совокупностью канонических OWL-подсхем.

На данном этапе имеется выделенная OWL-схема объектной надстройки над реляционным хранилищем данных. Для возможности интеграции информационных ресурсов репозиториев различных внешних систем, описанных подобными схемами, необходимо выделить из них канонические (общие) подсхемы, в рамках которых будут формироваться объектные запросы на доступ к информационным ресурсам и осуществляться интеграция полученных от различных внешних систем ответов. В свете этого на данном этапе пространство имен прикладной OWL-структуры разбивается на следующие три:

– пространство имен common – каноническая OWL-подструктура общих классов, свойств, в соответствии с которыми могут формироваться объектные запросы;

– пространство имен external – каноническая OWL-подструктура общих прикладных классов, свойств, в соответствии с которыми пользователю могут возвращаться данные прикладной системы;

– пространство имен external_own – OWL-подструктура общих прикладных классов, свойств, которые поддерживаются репозиторием, но недоступны объектным запросам.

6. Реализация адаптера объектного репозитория – поддержка прикладных OWL-структур, согласованных с совокупностью канонических OWL-структур. На данном этапе сформированное полноценное описание объектной схемы данных репозитория используется как входной параметр для реализованного адаптера объектного репозитория, который позволяет:

– осуществить объектно-реляционное отображение полученной объектной схемы данных на реляционную схему существующей реляционной БД;

– выполнять объектные OQL-запросы к репозиторию, согласованные с канонической OWL-структурой общих классов;

– представлять результаты OQL-запросов к репозиторию в унифицированном OWL/XML формате;

– предоставить Web-сервис для выполнения OQL-запросов к сформированному объектному репозиторию и получения OWL/XML ответов.

Выводы

Предлагаемая модель представления и обработки данных в Интернет на основе семантических сетей позволит в значительной степени повысить описательную способность метаданных для извлечения и обработки информации. Она может

применяться как для формирования интеллектуальных сервисов поверх существующих реляционных баз данных, так и для построения различных интеллектуальных систем. Пользователи таких систем не только осуществляют доступ к потенциально более качественному сервису по сравнению с традиционным подходом, но и получают возможность активно участвовать в моделировании запросов, возможно, с применением визуального или естественно-языкового интерфейса. Предложенная методика перехода от реляционной модели данных к семантической позволяет рассматривать информацию, представленную в реляционной базе данных, в терминах теории искусственного интеллекта, что делает возможным интеллектуальную обработку данных.

Литература

1. Кодд Э. Расширение реляционной модели для лучшего отражения семантики // СУБД. – 1996. – № 5-6.
2. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб.: Питер, 2001.
3. Цаленко М.Ш. Моделирование семантики в базах данных. – М.: Наука, 1989.
4. Некрашевич С.П. Агентно-ориентированный подход к разработке сложных программных систем // Харьков: Вестник НТУ«ХПИ». – 2004. – № 36. – 178 с.
5. Кузнецов С.Д. Направления исследований в области управления базами данных: краткий обзор // СУБД. – 1995. – № 1.

С.П. Некрашевич, Д.В. Божко

Представлення даних в Інтернет на основі семантичних мереж

У статті розглядаються сучасні моделі представлення та обробки даних в Інтернет. Проаналізовані реляційна модель представлення даних та деякі моделі представлення знань. Запропоновано засоби та методи підвищення семантичних відношень в інформації, на основі котрих відбувається пошук в мережі Інтернет. Приведено опис алгоритму переходу від реляційної моделі даних до семантичної мережі. Описано можливі способи застосування розробленої моделі.

S.P. Nekrashevich, D.V. Bozhko

Data Presentation in Internet on Basis of Semantic Net

In the given article the modern models of data presentation and data processing in Internet are examined. The relational model of data presentation and some models of knowledge presentation are analyzed. The means and methods to increase the semantic relations of information are offered for using in the Internet searching. The description of the algorithm to convert relational model to semantic net is given. The possible usage of the developed model is described.

Статья поступила в редакцию 13.07.2005.