Концепция и архитектура тематического интеллектуального научного интернет-ресурса

© Ю. А. Загорулько © Г. Б. Загорулько © В. К. Шестаков © И. С. Кононенко
Институт систем информатики имени А.П. Ершова СО РАН,
Новосибирск
zagor@iis.nsk.su gal@iis.nsk.su umanist@gmail.com irina_k@cn.ru

Аннотация

В докладе представлен подход к обеспечению содержательного доступа к научным информационным ресурсам, представляющим определенную область знаний, и средствам их интеллектуальной обработки путем создания тематических интеллектуальных научных интернет-ресурсов (ИНИР). Тематический ИНИР представляет собой информационную систему, базирующуюся на формализме онтологий и семантических сетей. Онтология наряду с описанием моделируемой области знаний содержит описание структуры и типологии интегрируемых информационных ресурсов и методов интеллектуальной обработки содержащихся в них данных и знаний. Семантическая сеть играет роль интеллектуального хранилища данных, в котором накапливается информация о релевантных научных информационных ресурсах и сервисах, реализующих методы интеллектуальной обработки информации. На основе онтологии и семантической сети организуется удобная навигация по научным знаниям и информационным ресурсам, интегрированным в ИНИР, а также содержательный поиск требующихся данных и средств их интеллектуальной обработки. Работа выполнена при финансовой поддержке РФФИ (проект № 13-07-00422) и Президиума РАН (интеграционный проект СО РАН № 15/10).

1 Введение

Несмотря на то, что в настоящее время накоплено громадное количество информации по различным областям знаний и при этом значительная ее часть представлена непосредственно в сети Интернет, проблема эффективного обеспечения научного сообщества информацией по интересующим его тематикам все еще остается не решенной. Такое положение объясняется тем, что современные информационные системы используют довольно ограниченный набор методов поиска, представления, интерпретации и обработки информации.

Такие системы в основном представляют и выдают пользователю знания и данные в виде отдельных текстовых документов, в то время как для человека наиболее естественной формой подачи информации является представление ее в виде сети взаимосвязанных фактов. При этом большая часть информации, представленной в Интернет и локальных хранилищах данных (электронных библиотеках, архивах и т.п.), становится практически недоступной из-за неэффективной работы поисковых машин, которые в основном применяют примитивные механизмы поиска по ключевым словам, не учитывающие ни семантику слов, входящих в запрос, ни его контекст.

Нерешенной остается и проблема удобного доступа к средствам обработки интересующей пользователя информации. Даже уже представленные в Интернет реализации методов обработки информации остаются недоступными широкому кругу пользователей из-за отсутствия содержательной информации о них.

Для представления информации о некоторой области знаний и содержательного описания релевантных ей информационных ресурсов и методов интеллектуальной обработки содержащихся в них данных и знаний удобным средством являются онтологии [8, 18] и семантические сети [21, 23]. Именно эти формализмы предлагается положить в основу информационных систем, предназначенных для осуществления информационной поддержки научной и производственной деятельности в определенной области знаний. Такие системы мы будем называть тематическими интеллектуальными научными интернет-ресурсами (ИНИР).

В докладе описывается подход к построению тематических ИНИР, который развивает подход к построению порталов научных знаний [19], успешно

Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2013, Ярославль, Россия, 14-17 октября 2013 г.

применявшийся при создании порталов знаний для ряда научных дисциплин [20]. Вторая глава посвящена описанию концепции ИНИР, в третьей главе представляется система знаний ИНИР, в четвертой – архитектура ИНИР.

2 Концепция интеллектуального научного интернет-ресурса

В соответствии с предлагаемой концепцией тематический ИНИР представляет собой доступную через Интернет информационную систему, обеспечивающую систематизацию и интеграцию научных знаний и информационных ресурсов определенной области знаний, содержательный эффективный доступ к ним (поиск и навигацию) и поддерживающую их использование при решении различных научных и производственных задач за счет предоставления соответствующих интерфейсов и сервисов.

Как было сказано выше, ИНИР базируется на формализмах онтологий и семантических сетей. При этом онтология составляет ядро информационной модели ИНИР и наряду с описанием моделируемой области знаний содержит соотнесенное с ним описание структуры и типологии интегрируемых информационных ресурсов и методов интеллектуальной обработки содержащихся в них данных и знаний.

Семантическая сеть, структура которой определяется онтологией ИНИР, играет роль интеллектуального хранилища данных, в котором накапливается информация о релевантных научных информационных ресурсах и web-сервисах, реализующих методы обработки содержащейся в них информации.

На основе онтологии и семантической сети организуется удобная навигация по научным знаниям и информационным ресурсам, интегрированным в ИНИР, а также содержательный поиск требующихся данных и средств их интеллектуальной обработки (представленных, в том числе, в виде web-сервисов).

Особенностью данного подхода является то, что ИНИР позволяет значительно сократить время, которое требуется для обеспечения доступа к необходимой информации и ее анализа, за счет предварительного поиска релевантных интернет-ресурсов и аккумуляции их описаний непосредственно в семантической сети ИНИР. С этой целью в программную оболочку ИНИР включается подсистема сбора онтологической информации (метаданных) о релевантных интернет-ресурсах, которая в своей работе опирается на онтологию и тезаурус области знаний ИНИР.

3 Система знаний интеллектуального научного интернет-ресурса

Формально система знаний ИНИР описывается четверкой:

$KS = \langle O, Th, SN, IRs \rangle$, где

O – онтология ИНИР,

Th – тезаурус области знаний ИНИР,

SN – семантическая сеть, служащая для представления информационного содержания (контента) ИНИР,

IRs – информационные ресурсы, интегрируемые в ИНИР, и средства их интеллектуальной обработки (web-сервисы).

Рис.1. Система знаний ИНИР

Как было сказано выше, онтология ИНИР кроме описания понятий и отношений моделируемой области знаний включает соотнесенное с ним описание структуры и типологии интегрируемых информационных ресурсов и методов интеллектуальной обработки содержащихся в них данных и знаний. В связи с этим онтология состоит из трех взаимосвязанных онтологий, отвечающих за представление указанных выше компонентов знаний, а именно: онтологии области знаний ИНИР, онтологии научных интернет-ресурсов и онтологии задач и методов.

Онтология области знаний ИНИР строится на основе двух базовых онтологий, включенных в программную оболочку ИНИР, – онтологии научной деятельности и онтологии научного знания.

Онтология научной деятельности базируется на онтологии, предложенной в [2] для описания научно-исследовательских проектов и расширенной в [19] для применения к более широкому классу задач. Эта онтология включает классы понятий, относящиеся к организации научной и исследовательской деятельности, такие как *Персона, Организация, Событие, Научная деятельность, Проект, Публикация* и др.

Онтология научного знания фиксирует основные содержательные структуры, используемые для построения онтологий областей знаний. В

частности, эта онтология содержит классы, задающие структуры для описания понятий конкретных областей знаний, такие как *Раздел науки, Метод исследования, Объект исследования, Научный результат* и др. (Подробное описание этой онтологии можно найти в [19].)

Классы рассмотренных выше базовых онтологий связаны между собой ассоциативными отношениями, выбор которых осуществлялся не только исходя из полноты представления области знаний ИНИР, но и с учетом удобства навигации по его информационному пространству и поиска информации.

Основным классом онтологии научных интернет-ресурсов является класс Информационный ресурс, который служит для описания, представленных в сети Интернет информационных ресурсов. Набор атрибутов и связей этого класса основан на стандарте Dublin core [12]. Его атрибутами являются: название ресурса, язык ресурса, тематика ресурса, тип доступа к ресурсу и т.п. Объекты этого класса могут быть связаны семантическими отношениями с другими информационными объектами, представляющими в контенте ИНИР организации, персоны, публикации, события, разделы науки и т.д.

Онтология задач и методов кроме описания задач, на решение которых нацелен ИНИР, и методов их решения включает также описания web-сервисов, реализующих методы обработки информации, содержащейся в интегрируемых в ИНИР информационных ресурсах. Описания таких web-сервисов базируются на онтологии OWL-S [16], предназначенной для описания семантических web-сервисов (Semantic Web Services) [11] и представленной на языке OWL [1].

Обычный web-сервис представляет собой программную систему, достаточно понятное описание функциональности которой представлено в сети Интернет в виде интерфейса, заданного в пригодном для машинной обработки формате [4]. Благодаря этому описанию web-сервис может быть найден другими программами, которые могут взаимодействовать с ним в соответствии с его описанием путем посылки XML-сообщений, передаваемых с помощью HTTP-протокола.

Для реализации web-сервисов существует много технологий, но наиболее распространенными из них являются две: SOAP и REST.

SOAP (раньше расшифровывалось как Simple Object Access Protocol) – это протокол обмена сообщениям в формате XML, чаще всего посредством протокола HTTP [9]. Для представления в Интернет сервисов данного типа используется спецификация, заданная в пригодном для машинной обработки формате на языке WSDL (Web Services Description Language) [3].

REST (Representational State Transfer) — это архитектурный стиль построения распределенных приложений [6]. Передача данных происходит в одном из стандартных форматов (кроме XML может использоваться, например, JSON [13]). В качестве сетевого протокола также, как правило, используется HTTP. В качестве спецификации возможно использование либо WSDL 2.0 [14], либо WADL (Web Application Description Language) [10].

У каждой из этих технологий есть своя область применения и выбор зависит от задачи. В частности, SOAP обеспечивает строгую типизацию, но является более громоздким, чем REST.

Отличие семантического web-сервиса от обычного web-сервиса состоит в том, что он предоставляет потенциальным пользователям не только описание своего интерфейса в терминах типов входных и выходных данных, но и описание своей семантики, т.е. того, что сервис делает, его предметной области, ограничений на область применения и качество сервиса и т.п. Причем все его свойства, «способности» (функциональность) и интерфейсы кодируются в однозначной поддающейся машиной обработке форме. Наличие семантического описания у таких сервисов обеспечивает не только реализацию их поиска и корректного использования (исполнения), но и возможность композиции из них новых сервисов с целью получения функциональности, требуемой для решения пользовательских задач.

Наличие у семантических web-сервисов содержательных описаний создает предпосылки и для их успешной интеграции в ИНИР. При этом будет обеспечиваться содержательный доступ к ним не только для программных агентов, но и для человека, желающего найти необходимые для решения его задач средства интеллектуальной обработки информации.

Возможна также интеграция в ИНИР и обычных (не семантических) web-сервисов. Для этого такие сервисы предварительно снабжаются семантическими описаниями, составленными с использованием понятий онтологии OWL-S и онтологии области знаний ИНИР, в соответствии с правилами, предложенными в [15].

Таким образом, онтология ИНИР O, вводя формальные описания понятий некоторой области знаний, типов интегрируемых информационных ресурсов и методов их интеллектуальной обработки в виде классов объектов и отношений между ними, одновременно задает структуры для представления информации о реальных объектах моделируемой области знаний, интегрируемых информационных ресурсах IRs и методах и средствах обработки содержащейся в них информации. Данная информация хранится в контенте ИНИР в виде объектно-ориентированной семантической сети SN, типы информационных объектов и отношений которой определяются классами объектов и отношений, введенных в онтологии ИНИР.

Если главным назначением онтологии ИНИР является формализация моделируемой области

знаний, то тезаурус Th служит для описания смысла понятий (терминов), используемых в этой области знаний. При этом тезаурус позволяет задавать смысл понятий не только с помощью определений, но и посредством соотнесения их с другими понятиями, используя для этого не только семантические отношения, задающие иерархии понятий, но и отношения синонимии и ассоциации. Благодаря этому тезаурус может применяться как при обработке пользовательских запросов, так и при поиске и аннотировании информационных ресурсов [22], интегрируемых в ИНИР.

Тезаурус ИНИР строится на основе ядра тезауруса, изначально включенного в систему знаний ИНИР. Ядро тезауруса содержит описание понятий базовых онтологий, включая описание терминов, с помощью которых они представляются в интернет-ресурсах.

Таким образом, система знаний ИНИР не только включает формальное описание области знаний ИНИР, задает типологию релевантных ей информационных ресурсов и решаемых в ней задач и методов их решения, описывает смысл используемых в этой области знаний понятий, но и обеспечивает эффективное представление информации об интегрируемых в ИНИР информационных ресурсах и средствах их интеллектуальной обработки.

4 Архитектура ИНИР

ИНИР имеет традиционную трехуровневую архитектуру (см. рис. 2), включающую уровень доступа к информации, уровень обработки информации и базовый уровень.

Первый уровень обеспечивается пользовательским интерфейсом – программным компонентом, предоставляющим конечному пользователю содержательный доступ к контенту ИНИР при решении его задач. Главными функциями этого интерфейса являются представление пользовательских запросов и результатов поиска и решений задач, а также обеспечение навигации в информационном пространстве ИНИР. Благодаря использованию онтологии и тезауруса пользовательский интерфейс позволяет формулировать запросы в терминах моделируемой области знаний и поддерживает управляемую онтологией навигацию по информационному пространству ИНИР.

Уровень обработки информации обеспечивает все информационные потоки в ИНИР – от конструирования онтологии до обработки пользовательских запросов. Он включает модуль поиска информации в информационном пространстве ИНИР, средства разработки/настройки базы знаний ИНИР и управления его контентом, а также подсистему сбора онтологической информации (метаданных) об интернет-ресурсах.

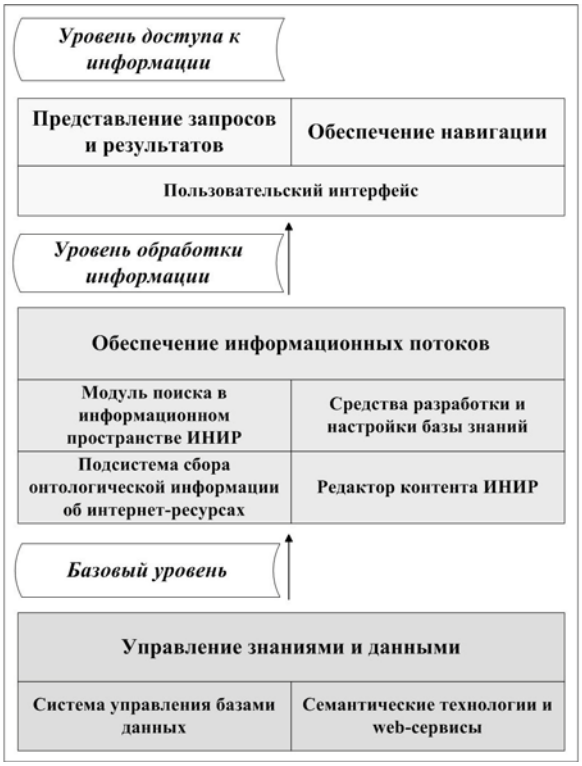

Рис. 2. Архитектура ИНИР

Настройка базы знаний выполняется с помощью средств разработки и настройки базы знаний, в состав которых входят редакторы онтологий и тезаурусов. Для управления контентом служит редактор данных, который позволяет создавать, редактировать и удалять информационные объекты и связи между ними. Все редакторы реализованы как web-приложения, поэтому обеспечивают удаленную настройку ИНИР и поддержку его контента авторизованными экспертами через Интернет.

Подсистема сбора онтологической информации об интернет-ресурсах выполняет поиск в Интернет информационных ресурсов, релевантных области знаний ИНИР, их семантический анализ и семантическое аннотирование на основе онтологии и тезауруса. Кроме того, эта подсистема формирует по полученным аннотациям представление информации об интернет-ресурсах в контенте ИНИР в виде информационных объектов класса Информационный ресурс и их связей с другими информационными объектами. Благодаря этому объекты, соответствующие проаннотированным информационным ресурсам, могут использоваться при поиске информации и навигации по контенту ИНИР.

Базовый уровень обеспечивает управление всеми данными и знаниями информационного пространства ИНИР. Он выполняет функции хранения и манипулирования данными (контентом ИНИР) и знаниями (онтологией и тезаурусом) с использованием средств стандартных СУБД (MySQL), технологий Semantic Web (OWL, RDF) и семантических web-сервисов (WSDL, OWL-S).

В частности, в качестве хранилища данных используется свободно распространяемая версия системы Virtuoso (Virtuoso Open-Source Edition) [17], обеспечивающая эффективную работу с большими объемами данных в RDF-формате (RDF Triple Store) [5]. Кроме того, эта система предоставляет возможность взаимодействия с web-сервисами.

5 Заключение

В докладе представлен подход к построению интеллектуальных научных интернет-ресурсов, обеспечивающих систематизацию и интеграцию информационных ресурсов определенной области знаний и средств интеллектуальной обработки представленной в них информации, а также содержательный эффективный доступ к ним и их использование при решении научных и производственных задач. Основу таких ИНИР составляет онтология, которая содержит наряду с описанием понятий и отношений моделируемой области знаний, соотнесенное с ним описание типологии интегрируемых информационных ресурсов и методов интеллектуальной обработки содержащихся в них данных и знаний. Семантическая сеть, структура которой определяется онтологией, играет в ИНИР роль интеллектуального хранилища данных, в котором накапливается информация о релевантных научных информационных ресурсах и семантических web-сервисах, реализующих методы интеллектуальной обработки содержащейся в них информации.

Данный подход развивает технологию разработки порталов научных знаний [19], успешно применявшуюся при создании порталов знаний для ряда научных дисциплин [20]. В отличие от порталов научных знаний ИНИР обеспечивает доступ не только к релевантным информационным ресурсам, но и средствам их интеллектуальной обработки (анализа), представленным в виде web-сервисов. Таким образом, ИНИР объединяет преимущества и порталов знаний, и мэшапов [7]. От первых он наследует представление знаний и данных в виде онтологии и семантической сети, от вторых – объединение различных web-сервисов в одном приложении.

Важным преимуществом предложенного подхода является также то, что ИНИР позволяет значительно упростить и ускорить доступ к затребованной пользователем информации и сократить время ее анализа благодаря аккумуляции описаний релевантных интернет-ресурсов и методов их обработки непосредственно в семантической сети ИНИР.

Литература

[1] Antoniou G., Harmelen F. Web Ontology Language: OWL // Handbook on Ontologies. – Berlin: Springer Verlag. – 2003. – p. 67-92.

[2] Benjamins V.R. and Fensel D. Community is Knowledge! in (KA)2 // Proceedings of the 11th Banff Knowledge Acquisition for Knowledge-based Systems workshop, KAW'98 (Banff, Canada, April 1998). – Calgary: SRDG Publications, Department of Computer Science, University of Calgary, 1998.

[3] Chinnici R., Moreau J.-J., Ryman A., Weerawarana S. Web Services Description Language (WSDL) Version 2.0 Part 1: Core Language. http://www.w3.org/TR/wsdl20/

[4] Erl T. Service-Oriented Architecture: Concepts, Technology, and Design. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2005. 792 p.

[5] Erling O., Mikhailov I. RDF Support in the Virtuoso DBMS // Networked Knowledge-Networked Media. – Springer Berlin Heidelberg, 2009. – p. 7-24.

[6] Fielding R. Architectural Styles and the Design of Network-based Software Architectures. http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

[7] Gheorghiu C., Nicolescu R., Bogdan A. V., Ochisor C., Buraga S. C., and Alboaie L. Sigma - semantic government mash-up application: Using semantic web technologies to provide access to governmental data. In Proc. of the 2011 10th International Symposium on Parallel and Distributed Computing (ISPDC'11). Washington, DC, USA. IEEE Computer Society, 2011, p. 247–253.

[8] Guarino N. Formal Ontology in Information Systems // Formal Ontology in Information Systems. Proceedings of FOIS'98, Trento, Italy, June 6–8, 1998 / Ed. N. Guarino. Amsterdam: IOS Press, 1998. p. 3–15.

[9] Gudgin M., Hadley M., Mendelsohn N., Moreau J.-J., Nielsen H.F., Karmarkar A., Lafon Y. SOAP Version 1.2 Part 1: Messaging Framework (Second Edition). http://www.w3.org/TR/soap12-part1/

[10] Hadley M. Web Application Description Language http://www.w3.org/Submission/wadl/

[11] Farshad Hakimpour, Suo Cong, Daniela E. Damm. A Practical Tutorial on Semantic Web Services. http://www.academia.edu/331267/A_Practical_Tutorial_on_Semantic_Web_Services

[12] Hillmann D. Using Dublin Core. http://dublincore.org/documents/usageguide/

[13] Introducing JSON. http://json.org

[14] Mandel L. Describe REST Web services with WSDL 2.0. http://www.ibm.com/developerworks/webservices/library/ws-restwsdl/

[15] David L. Martin, Massimo Paolucci, Sheila A. McIlraith, Mark H. Burstein, Drew V. McDermott, Deborah L. McGuinness, Bijan Parsia, Terry R. Payne, Marta Sabou, Monika Solanki, Naveen Srinivasan, Katia P. Sycara. Bringing Semantics to Web Services: The OWL-S Approach. In proceeding of: Semantic Web Services and Web Process Composition, First International

Workshop, SWSWPC 2004, San Diego, CA, USA, July 6, 2004, Revised Selected Papers. Cardoso and A. Sheth (Eds.): SWSWPC 2004. LNCS 3387, p. 26–42. Berlin Heidelberg: Springer-Verlag, 2005.

[16] OWL-S: Semantic Markup for Web Services. http://www.w3.org/Submission/OWL-S/

[17] Virtuoso open-source edition. OpenLink Software. http://www.openlinksw.com/wiki/main

[18] Гаврилова Т.А., Кудрявцев Д.В., Горовой В.А. Модели и методы формирования онтологий // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. 2006. № 46. с. 21-28

[19] Загорулько Ю.А., Боровикова О.И. Подход к построению порталов научных знаний // Автометрия. № 1, 2008, т. 44. –с. 100–110.

[20] Загорулько Ю.А., Боровикова О.И., Загорулько Г.Б. О применении технологии создания порталов научных знаний // Тр. XV Байкальской Всероссийской конф. «Информационные и математические технологии в науке и управлении». – Иркутск: Институт систем энергетики им Л.А. Мелентьева СО РАН, 2010. –т.2. –с. 164–171.

[21] Лозовский, В.С. Семантические сети // Представление знаний в человеко-машинных и робототехнических системах. – М.: ВИНИТИ, 1984. –Т.А. – с. 84-121.

[22] Лукашевич Н.В. Тезаурусы в задачах информационного поиска. –М.: Изд-во МГУ, 2011.

[23] Осипов Г.С. Построение моделей предметных областей. Неоднородные семантические сети // Известия АН СССР. Техническая кибернетика. –1990. – №5. – с. 32–45.

Concept and architecture of thematic intelligent scientific Internet resource

Yury A. Zagorulko, Galina B. Zagorulko,

Vladimir K. Shestakov, Irina S. Kononenko

The paper presents an approach to providing content-based access to scientific information resources representing certain area of knowledge and to facilities for intelligent processing of information contained in these resources by means of creation of thematic intelligent scientific Internet resources (ISIR). The thematic ISIR is an Internet accessible information system which is based on ontology and semantic network formalisms. The ontology is a basis of INIR knowledge system. It includes along with descriptions of the modeled area of knowledge also descriptions of structure and typology of the integrated information resources relating to this area and methods of intelligent processing of knowledge and data contained in them. The semantic network whose structure is defined by the ontology plays role of an intelligent data warehouse where information about scientific information resources and web-services implementing methods of intelligent processing of information contained in these resources is accumulated. Based on the ontology and semantic network both a convenient navigation through scientific knowledge and information resources integrated in the INIR and content-based search of required data and facilities for their intelligent processing are organized.