УДК 002.53:004.89

Сбор онтологической информации для интеллектуальных научных Интернет-ресурсов

Загорулько Ю.А. (Институт систем информатики СО РАН),

Боровикова О.И. (Институт систем информатики СО РАН),

Сидорова Е.А. (Институт систем информатики СО РАН),

Ахмадеева И.Р. (Новосибирский государственный университет)

В работе рассматриваются проблемы сбора информации для тематических интеллектуальных научных интернет-ресурсов, обеспечивающих систематизацию и интеграцию научных знаний, информационных ресурсов, относящихся к определенной области знаний, и средств их интеллектуальной обработки, а также содержательный доступ к ним. Предлагается подход к автоматизации сбора информации, объединяющий методы метапоиска и извлечения информации, базирующиеся на онтологиях и тезаурусах моделируемой области знаний.

Ключевые слова: *научный интернет-ресурс, метапоиск, извлечение информации, онтология, тезаурус.*

1. Введение

В мире накоплено огромное количество информации по различным областям знаний, причем значительная ее часть представлена непосредственно в сети Интернет, но, несмотря на это, проблема эффективного обеспечения научного сообщества информацией по интересующим его тематикам остается на повестке дня. Нерешенной остается и проблема удобного доступа к средствам обработки данных, собранных по этим тематикам. Даже уже представленные в Интернет в виде веб-сервисов реализации методов их обработки остаются недоступными широкому кругу пользователей из-за отсутствия содержательной информации о них.

Для решения этих проблем, в частности для информационной и аналитической поддержки научной и производственной деятельности в определенных областях знаний создаются тематические интеллектуальные научные интернет-ресурсы (ИНИР) [4].

Эффективность использования ИНИР будет тем выше, чем более полно в нем будет представлена информация по его тематике. Однако сбор и накопление такой информации – довольно трудоемкая задача, решить которую можно только за счет автоматизации сбора релевантной информации по тематике ИНИР из сети Интернет. Описанию подхода, поддерживающего такую автоматизацию, и посвящена данная работа.

Работа выполнена при финансовой поддержке РФФИ (проект № 13-07-00422) и Президиума РАН (интеграционный проект СО РАН № 15/10).

2. Информационная модель ИНИР

Тематический ИНИР представляет собой информационную систему, обеспечивающую систематизацию и интеграцию научных знаний и информационных ресурсов определенной области знаний, содержательный эффективный доступ к ним (поиск и навигацию) и средствам их интеллектуальной обработки.

Ядро информационной модели ИНИР составляет онтология, которая, вводя формальные описания понятий некоторой области знаний, типов 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 и методах и средствах их обработки. Данная информация хранится в контенте ИНИР в виде семантической сети, типы информационных объектов и отношений которой определяются классами объектов и отношений онтологии ИНИР.

На основе онтологии организуется удобная навигация по научным знаниям и информационным ресурсам, интегрированным в ИНИР, а также содержательный поиск данных и средств их интеллектуальной обработки.

В систему знаний ИНИР включен также тезаурус, содержащий термины моделируемой области знаний, т.е. слова и словосочетания, с помощью которых понятия онтологии представляются в текстах и пользовательских запросах. Тезаурус задает смысл понятий посредством соотнесения одних понятий с другими с помощью семантических отношений. Благодаря этому он может применяться при поиске и аннотировании информационных ресурсов, интегрируемых в ИНИР.

Поскольку предлагаемый подход к сбору информации из сети Интернет существенно базируется на онтологии ИНИР, рассмотрим ее подробнее.

Онтология ИНИР состоит из трех взаимосвязанных онтологий: онтологии области знаний ИНИР, онтологии научных информационных ресурсов и онтологии задач и методов.

Онтология области знаний ИНИР строится на основе двух базовых онтологий – онтологии научной деятельности и онтологии научного знания. Первая из этих онтологий включает классы понятий, относящиеся к организации научной и исследовательской деятельности, такие как *Персона, Организация, Событие, Конференция, Проект, Публикация* и др. Вторая базовая онтология содержит понятия, необходимые для представления научных дисциплин. В частности, она содержит такие классы, как *Раздел науки, Метод исследования, Объект исследования, Научный результат* и др.

Онтология задач и методов предназначена для описания задач, на решение которых нацелен ИНИР, и методов их решения. Кроме этих методов, в онтологии представлены методы интеллектуальной обработки данных, содержащихся в интегрируемых в ИНИР информационных ресурса, а также описания web-сервисов, реализующих эти методы.

Основным классом онтологии научных информационных ресурсов является класс *Информационный ресурс*, служащий для описания информационных ресурсов. Набор атрибутов и связей этого класса основан на стандарте Dublin core [8]. Он имеет следующие атрибуты: *название ресурса, язык ресурса, тематика ресурса, тип доступа к ресурсу* и т.п. Объекты этого класса связываются семантическими отношениями с другими информационными объектами, представляющими в контенте ИНИР организации, персоны, публикации, проекты, разделы науки и т.п.

3. Модель сбора информации

Прежде чем представить предлагаемую модель, заметим, что проблемой сбора информации из Интернет занимаются многие исследователи и разработчики. Однако, как показывает обзор [7], большая часть таких исследований направлена на извлечение информации, необходимой для решения задач электронной коммерции или анализа новостного потока и социальных сетей, и лишь их незначительная часть – на извлечение информации для нужд научной деятельности [5; 6].

Сложность задачи сбора информации для ИНИР определяется большим разнообразием видов извлекаемой информации и способов ее представления в Интернет. В частности, необходимо собирать информацию об организациях, проектах, публикациях, интернет-ресурсах, веб-сервисах и других сущностях, описываемых онтологией научной деятельности. Эта информация может быть представлена как в виде интернет-страниц, имеющих различную структуру, так и в виде текстовых документов в различных форматах. В связи с этим мы посчитали нецелесообразным использовать популярные в настоящее время методы извлечения информации, основанные на обучении на примерах (см. например, [10]), а

применили подход, базирующийся на онтологии. В соответствии с этим было решено для каждого типа сущностей (класса онтологии научной деятельности) разработать свой метод сбора и обработки информации, настраиваемый на предметную область и типы интернет-ресурсов и документов.

Заметим, что предлагаемый подход развивает методы сбора онтологической информации об интернет-ресурсах [2], разработанные в рамках технологии построения порталов научных знаний. В то же время он близок к подходу, представленному в [6], который базируется на концептуальной модели предметной области и предлагает использовать для каждого вида сущности свои шаблоны и обработчики.

Сбор информации для ИНИР включает следующие этапы:

1. Поиск релевантных области знаний ИНИР интернет-ресурсов и документов.
2. Извлечение информации из найденных интернет-ресурсов и документов.
3. Занесение полученной информации в контент ИНИР.

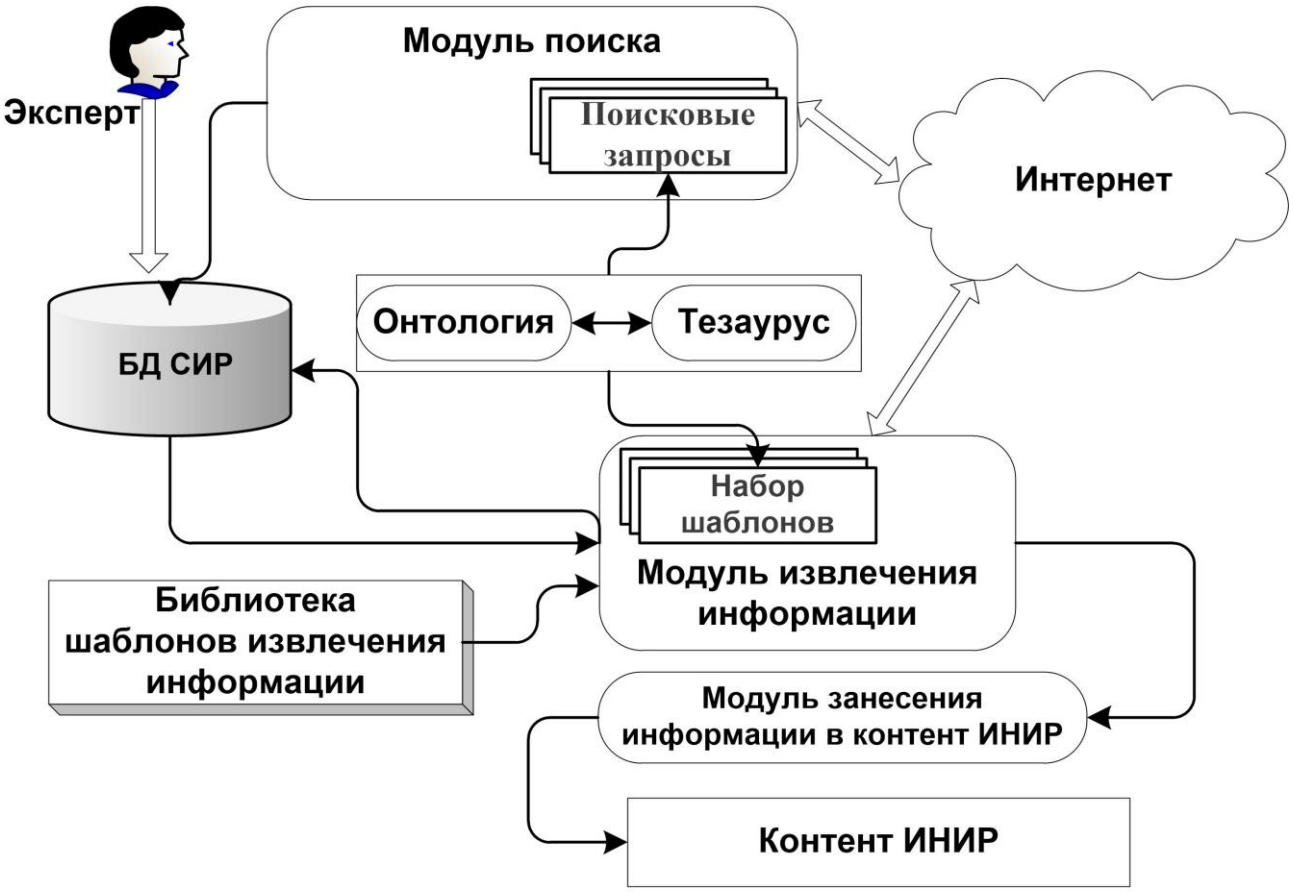

Рис.1. Подсистема сбора информации из Интернет

В соответствии с предложенной схемой подсистема сбора информации из сети Интернет включает следующие компоненты (см. Рис.1): модуль поиска релевантных интернет-

ресурсов, модуль извлечения информации из интернет-ресурсов, модуль занесения найденной информации в контент ИНИР, а также базу данных ссылок на интернет-ресурсы (БД СИР).

3.1. Сбор релевантных интернет-ресурсов

При настройке ИНИР на область знаний выполняется заполнение БД СИР ссылками на релевантные, по мнению экспертов, интернет-ресурсы. При этом для каждой ссылки указывается класс (классы) онтологии, объект (объекты) которого описывает соответствующий ей ресурс. С каждой ссылкой также связывается следующая метаинформация: дата загрузки, частота обновления, периодичность повторной закачки, дата последней проверки, статус обработки. Первые четыре параметра вводятся для отслеживания актуальности ресурса, последний – для указания статуса ссылки (релевантная, нерелевантная, необработанная).

Список ссылок пополняется не только вручную, но и автоматически – модулем поиска интернет-ресурсов, который выполняет сбор ссылок на релевантные интернет-ресурсы по поисковым запросам, сформированным на основе названий классов онтологии и терминов тезауруса, представляющих понятия моделируемой области знаний. Он запускается с заданной при настройке ИНИР периодичностью. При этом модуль поиска обращается к поисковым системам Google, Яндекс и Bing через их программные интерфейсы, т.е. использует механизм метапоиска с последующей фильтрацией дубликатов и нерелевантных ссылок [1].

БД СИР может также пополняться ссылками, обнаруженными при извлечении информации из обрабатываемых интернет-ресурсов. Эти ссылки в дальнейшем анализируются экспертами, которые принимают решение об их релевантности.

3.2. Методы извлечения информации

Для заполнения контента ИНИР собирается информация из таких источников, как порталы знаний, электронные библиотеки и журналы, сайты организаций, ассоциаций, проектов и конференций, новостные ленты, социальные научные сети, вики-ресурсы, реестры (каталоги) веб-сервисов и др. Как было сказано выше, из этих источников извлекается информация о проектах, организациях, персонах, конференциях и публикациях, т.е. обо всех объектах базовых классов онтологии научной деятельности, а также информация о самих источниках, которая представляется в контенте ИНИР в виде объектов классов онтологии научных информационных ресурсов.

Для каждого из этих классов создается свой метод извлечения информации, включающий набор шаблонов и связанных с ним обработчиков. Шаблоны генерируются на основе онтологии. Для повышения полноты извлечения информации вариативность этих шаблонов увеличивается за счет использования альтернативных терминов из тезауруса (синонимов и гипонимов). В шаблонах для каждого типа извлекаемой информации указываются обработчики, реализующие алгоритмы обхода и анализа соответствующих фрагментов интернет-страниц или документов.

Модуль извлечения информации осуществляет анализ интернет-ресурсов, которые он скачивает по ссылкам, заданным в БД СИР.

Для облегчения анализа HTML-страница ресурса представляется в виде DOM-дерева в соответствии со стандартом DOM (Document Object Model), регламентирующим способ представления содержимого документа (в частности, HTML-страницы) в виде набора объектов [9]. Анализ DOM-дерева каждой страницы выполняется на основе соответствующего шаблона, при этом определяется релевантность загруженной страницы тематике ИНИР и извлечение описанной этим шаблоном информации.

Например, интернет-ресурс, на котором размещена информация о проекте, может быть представлен сайтом проекта, разделом сайта организации или персоны или публикацией, описывающей проект. Для каждого из этих способов представления на основе класса онтологии *Проект* строится свой шаблон.

Рассмотрим один из вариантов организации такого шаблона, но сначала дадим описание свойств класса *Проект* онтологии научной деятельности (см. Рис.2).

```
class Проект (Название: string; Аббревиатура: string; Описание: string;
      Дата начала: date; Дата окончания: date; Номер: string; URL: string;
      Стадия: Этап_проекта; Ключевые слова: set_of_string)
relation Проект_Включает < Проект, Проект >
relation Проект_Поддерживается < Проект, Организация >
relation Задача_Проекта < Проект, Задача >
relation Участник_Проекта < Проект, Персона > (Роль: set_of_Роль)
relation Участник_Проекта_Орг < Проект, Организация >
relation Научное_направление < Проект, Раздел_науки >
relation Результат_Деятельности < Проект, Результат >
relation Исследует_Объект < Проект, Объект_исследования >
relation Использует_Метод < Проект, Метод_исследования >
relation Публикация_о_Проекте < Проект, Публикация >
relation Интернет_Ресурс_Проекта <Проект,Интернет_Ресурс>
```

Рис.2. Описание класса *Проек*т

На рис.3 представлен шаблон для извлечения информации на основе класса *Проект*.

```
<Class Name= "Проект" engine = " FragmentSearch " >
  <Marker Term = "О проекте" PType= "Menu/Head" FragType= "Page/Block " />
  <Marker Term = "Проект" PType= " Head " FragType= " Block " />
          <Attr Name= "Название" type= "string" engine = "NameEntity" >
             <Marker Term = "Проект" Ptype = "link" FragType= "LinkText" />
             <Marker Term = "Проект" Ptype = "sentence" FragType= "QuoteText" />
             <Marker Term = "Проект" Ptype = "Head" FragType = "Head" />
          </Attr>
          <Attr Name = "Аннотация" type= "text" >
             <Marker Term = "Аннотация/Содержание проекта/Описание
             проекта/ О проекте " PType = "Head" FragType= "Block/Page" />
           </Attr>
<Relation Name = "Публикация_о_Проекте" >
  <Marker Term = "Публикации" PType= "Menu/Head" FragType="Page/Block" />
  <Marker Term = "Список публикаций" PType="Menu" FragType="Page" />
  <Marker Term = "Литература" PType= "Menu/Head" FragType="Page/Block" />
  <Marker Term = "Библиография" PType= "Menu/Head" FragType= "Page/Block" />
<Object Name = "Публикация"engine ="PublicationsList" />
</Relation>
<Relation Name = "Участник проекта" >
  <Marker Term = "Об участниках" PType= "Menu" FragType="Page" />
  <Marker Term = "Список участников" PType= "Head" FragType="Block"/>
  <Marker Term = "Исполнители" PType= "Head" FragType="Block"/>
  <Marker Term = "Участники" PType= "Head" FragType="Block" />
<Object Name = "Персона" engine = "PersonList" />
</Relation>
</Class>
```

Рис.3. Шаблон класса *Проект*

Шаблон, предназначенный для извлечения объектов заданного класса, описывается блоком **Class** и содержит блоки атрибутов (**Attr**), отношений (**Relation**) и аргументов отношений (**Object**). Каждый из этих блоков может описываться одним или группой альтернативных маркеров (**Marker**), задающих свойства фрагмента текста, содержащего извлекаемую информацию. Маркер, приписанный непосредственно блоку **Class**, выделяет текстовый фрагмент, описывающий объект и определяющий область дальнейшего поиска маркеров.

К параметрам маркера относятся: (1) Term – термин тезауруса, представленный множеством альтернативных написаний термина, (2) PType – тип фрагмента, в тексте которого должен располагаться термин маркера, (3) FragType – тип фрагмента, который должен извлекаться, (4) engine – имя обработчика, который будет извлекать требуемую информацию в найденном по маркеру фрагменте.

Анализ входной страницы, представленной после предварительной обработки в структурированном виде (DOM-структура), осуществляется обработчиком верхнего уровня, который решает следующие задачи:

- поиск шаблона, подходящего для данной страницы или ее фрагмента, на основе маркеров блока **Class**;
- поиск маркерных терминов и извлечение текстовых фрагментов в соответствии с параметрами маркера;
- вызов специализированных обработчиков, формирование входных данных и обработка результата их работы;
- формирование объекта заданного онтологического класса и его связей.

Рис.4. Сайт проекта «Национальный корпус русского языка»

Например, на сайте проекта «Национальный корпус русского языка» (http://www.ruscorpora.ru) в разделе меню «о проекте» можно найти краткое описание

проекта, в разделе «участники проекта» – информацию о персонах и организациях, участвующих в проекте, в разделе «публикации» – информацию о публикациях по теме проекта и т.д. (Рис.4)

Шаблон, построенный на основе класса *Проект*, позволит извлечь эту информацию со страниц данного сайта. При этом для извлечения информации, составляющей контекст проекта и, как правило, определяемой отношениями класса *Проект*, например, данных о публикациях по теме проекта, персонах и организациях, участвующих в проекте, используются обработчики и шаблоны, специально построенные для извлечения информации такого типа и многократно используемые в других шаблонах, соответствующих таким базовым понятиям онтологии, как *Публикация*, *Персона*, *Организация* и др.

5. Заключение

Тематический интеллектуальный научный интернет-ресурс позволяет исследователям значительно сократить время, требуемое для обеспечения доступа к необходимой информации и ее анализа, за счет аккумуляции в своем контенте описаний релевантных интернет-ресурсов и методов их обработки. При этом использование ИНИР будет тем эффективнее, чем более полно в нем будет представлена информация по его тематике. Добиться такой полноты можно только за счет автоматизации сбора информации из сети Интернет.

В настоящее время реализован ряд компонентов подсистемы сбора информации из сети Интернет, а именно: модуль поиска релевантных интернет-ресурсов, модуль извлечения информации, база данных ссылок на интернет-ресурсы. На данный момент разработаны методы извлечения информации о проектах, организациях и событиях, включая сопутствующие шаблоны и обработчики, реализующие извлечение информации о персонах и публикациях. Заметим, что для анализа списков публикаций и персон используются ранее разработанные нами средства генерации формальных описаний научных статей [3].

Список литературы

1. Ахмадеева И.Р., Загорулько Ю.А., Саломатина Н.В., Серый А.С., Сидорова Е.А., Шестаков В.К. Подход к формированию тематических коллекций текстов на основе интернет-ресурсов // Вестник НГУ. Серия: Информационные технологии. 2013. Том.11, выпуск 4. С. 5-15.
2. Загорулько Ю.А. Автоматизация сбора онтологической информации об интернет-ресурсах для портала научных знаний // Известия Томского политехнического университета. Т. 312. № 5. Управление, вычислительная техника и информатика. 2008. С. 114–119.

3. Загорулько Ю.А., Дяченко О.О. Автоматическое наполнение информационных систем библиографическими сведениями о научных публикациях // Труды XIII Всероссийской научной конференции RCDL'2011 «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Воронеж, 19-22 октября 2011 г. Воронеж: Издательско-полиграфический центр Воронежского государственного университета, 2011 С.347-353.
4. Загорулько Ю.А., Загорулько Г. Б., Шестаков В.К., Кононенко И.С. Концепция и архитектура тематического интеллектуального научного интернет-ресурса // Труды XV Всероссийской научной конференции RCDL'2013. 14-17 октября 2013 г. Ярославль: ЯрГУ, 2013. С.57–62.
5. Ланин В.В., Мальцев П.А., Лядова Л.Н. Технологии сбора и анализа информации для исследовательского портала // Материалы Четвертой международной научно-технической конференции «Инфокоммуникационные технологии в науке, производстве и образовании» (Инфоком 4): Часть I, 2010. С. 218–222.
6. DeRose P., Shen W., Chen F., Doan AH, Ramakrishnan R. Building Structured Web Community Portals: A Top-Down, Compositional, and Incremental Approach // VLDB '07, September 23-28, 2007, Vienna, Austria. P. 399-410.
7. Ferrara E., De Meo P., Fiumara G., Baumgartner R.. Web Data Extraction, Applications and Techniques: A Survey // Preprint submitted to Knowledge-based systems. June 5, 2014. 41p.
8. Hillmann D. Using Dublin Core, 2005. [Электронный ресурс]. URL: http://dublincore.org/documents/usageguide/ (дата обращения: 17.11.2014).
9. Stenback J., Le Hégaret P., Le Hors A. Document Object Model (DOM) Level 2 HTML Specification // W3C Recommendation, 2003. [Электронный ресурс]. URL: http://www.w3.org/TR/2003/REC-DOM-Level-2-HTML-20030109/ (дата обращения: 17.11.2014).
10. Zhai Y., Liu B. Extracting Web Data Using Instance-Based Learning // Proceedings of 6th International Conference on Web Information Systems Engineering (WISE-05), 2005. P. 318–331.

UDK 002.53:004.89

An ontological information collection for intelligent scientific internet resources

Yury A. Zagorulko (A.P. Ershov Institute of Informatics Systems),

Olesya I. Borovikova (A.P. Ershov Institute of Informatics Systems),

Elena A. Sidorova (A.P. Ershov Institute of Informatics Systems),

Irina R. Ahmadeeva (Novosibirsk State University)

The paper considers the problems of information collection for thematic intelligent scientific internet resources providing the systematization and integration of scientific knowledge, information resources, related to certain area of knowledge, and methods of intelligent processing of data contained in them as well as the content-based access to them. The approach to automatization of information collection combining metasearch and knowledge extraction methods based on using ontology and thesaurus of the modeled area of knowledge is proposed.

Keywords: *Scientific Internet resources, metasearch, information extraction, ontology, thesaurus.*