Автоматизация сбора информации о научной деятельности для тематических интеллектуальных научных интернет-ресурсов

© Ю. А. Загорулько © И. Р. Ахмадеева © А. С. Серый
Институт систем информатики имени А.П. Ершова СО РАН,
Новосибирск
zagor@iis.nsk.su ah.irishka@gmail.com Alexey.Seryj@iis.nsk.su

Аннотация

В работе рассматриваются проблемы сбора информации для тематических интеллектуальных научных интернет-ресурсов, обеспечивающих систематизацию и интеграцию научных знаний, информационных ресурсов и средств интеллектуальной обработки информации, относящихся к определенной области знаний, а также содержательный доступ к ним. Предлагается подход к автоматизации сбора информации о научной деятельности в заданной области знаний, объединяющий методы метапоиска и извлечения информации, базирующиеся на онтологиях и тезаурусах. В соответствии с этим подходом для каждого типа сущностей (класса онтологии) разрабатываются свои методы сбора и извлечения информации, настраиваемые на область знаний и типы интернет-ресурсов.
Работа выполнена при финансовой поддержке РФФИ (проект № 13-07-00422).

1 Введение

Проблема обеспечения эффективного доступа к знаниям, произведенным в интересующих исследователя областях знаний, а также к информации о выполняемой в них научной деятельности и полученных результатах не имеет пока удовлетворительного решения.

Для решения данной проблемы была предложена концепция тематического интеллектуального научного интернет-ресурса (ИНИР) [12], обеспечивающего информационную и аналитическую поддержку научной деятельности в определенной области знаний.

Эффективность использования такого 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 из сети Интернет. Описанию методов, на которых базируется такая автоматизация, посвящена данная работа.

Следует заметить, что проблемой сбора информации из Интернет занимаются многие исследователи. Однако, как показывает обзор [3], большая часть таких исследований направлена на извлечение информации, необходимой для решения задач электронной коммерции или анализа новостного потока и социальных сетей. Что касается извлечения информации для нужд научной деятельности, то львиная доля таких работ посвящена извлечению информации из научных статей [1, 13], которые в основном содержат теоретические сведения об определенных областях знаний, а вот извлечению информации о выполняемой в этих областях научной деятельности, которая большей частью представлена в Интернет в виде веб-сайтов или веб-страниц, не уделяется должного внимания.

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.

Ядром системы знаний ИНИР является онтология (ONT) [4], которая, вводя формальные описания понятий некоторой области знаний, типов информационных ресурсов и методов их интеллектуальной обработки в виде классов объектов C и отношений между ними R, 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 объектно-ориентированной семантической сети, типы информационных объектов и отношений

Труды XVII Международной конференции DAMDID/RCDL'2015 «Аналитика и управление данными в областях с интенсивным использованием данных», Обнинск, 13-16 октября 2015

которой определяются классами объектов и отношений онтологии ИНИР.

Объектно-ориентированная семантическая сеть представляется двойкой $SN = < I_C, I_R >$,

где $I_C = < I_{C1}, ..., I_{Cn} >$ – множество объектов, т.е. экземпляров классов C, определенных в онтологии ONT, т.е. $\forall i : I_{Ci} \in C_i, C_i \in C$,

$I_R = < I_{R1}, ..., I_{Rm} >$ – множество экземпляров отношений R, определенных в онтологии ONT и связывающих объекты из множества I_C.

Онтология ИНИР состоит из трех взаимосвязанных онтологий, отвечающих за представление указанных выше трех компонентов знаний, а именно: онтологии области знаний ИНИР, онтологии научных интернет-ресурсов и онтологии задач и методов.

Онтология области знаний задает систему понятий и отношений, предназначенных для детального описания области знаний ИНИР и выполняемой в ее рамках научной и исследовательской деятельности. В частности, она содержит классы *Раздел науки*, *Метод исследования*, *Объект исследования*, *Научный результат*, используя которые, можно описать значимые для моделируемой области знаний разделы и подразделы, задать типизацию методов и объектов исследования, описать результаты научной деятельности. Для представления научной и исследовательской деятельности в онтологии служат классы *Персона*, *Организация*, *Событие*, *Научная деятельность*, *Проект*, *Публикация* и др.

Онтология научных интернет-ресурсов служит для описания, представленных в сети Интернет информационных ресурсов, релевантных области знаний ИНИР. Основным классом этой онтологии является класс *Информационный ресурс*, набор атрибутов и связей которого основан на стандарте Dublin core [5].

Онтология задач и методов кроме описания задач, для решения которых предназначен ИНИР, и методов их решения включает также описания web-сервисов, реализующих как эти методы, так и методы обработки информации, разработанные в моделируемой области знаний.

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).

Система знаний ИНИР также включает многоязычный тезаурус, содержащий термины моделируемой области знаний на нескольких языках (в настоящее время на русском и английском), 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.

3 Особенности подхода к сбору информации о научной деятельности

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 и других сущностях, описываемых онтологией ИНИР. 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, [9]), а применили подход, базирующийся на онтологии. В соответствии с этим было решено для каждого типа сущностей моделируемой области знаний разработать свой метод сбора и обработки информации, настраиваемый на эту область и типы интернет-ресурсов и документов.

Заметим, что предлагаемый подход развивает методы сбора онтологической информации об интернет-ресурсах [11], разработанные в рамках технологии построения порталов научных знаний. В то же время он близок к подходу, представленному в [2], который базируется на концептуальной модели предметной области и предлагает использовать для каждого вида сущности свои шаблоны и обработчики, а также подходам [6, 7], непосредственно базирующимся на онтологиях.

Сбор информации для ИНИР включает следующие этапы:

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

В соответствии с предложенной схемой подсистема сбора информации из сети Интернет включает следующие компоненты (см. Рис.1): модуль поиска, модуль извлечения информации, модуль занесения информации в контент ИНИР, а также базу данных ссылок на интернет-ресурсы (БД СИР).

Рис. 1. Схема подсистемы сбора информации

4 Сбор релевантных интернет-ресурсов

При настройке ИНИР на область знаний не только строится онтология ИНИР, но и заполняется БД СИР ссылками на релевантные интернет-ресурсы. С каждой ссылкой, помещаемой в БД СИР, связывается класс (классы) онтологии, объект (объекты) которого описывает соответствующий ей ресурс, а также метаинформация, необходимая для отслеживания актуальности и статуса релевантности ресурса.

Заметим, что заполнение БД СИР может выполняться как вручную – экспертами моделируемой области знаний, так и автоматически – модулем метапоиска.

Модуль метапоиска выполняет сбор ссылок на релевантные интернет-ресурсы по поисковым запросам, сформированным на основе названий классов онтологии и терминов тезауруса, представляющих понятия моделируемой области знаний. Такие запросы генерируются для всех языков, используемых в ИНИР. Метапоиск 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 [10].

Оценка релевантности ссылок выполняется с использованием характеристических векторов, которые строятся для запроса и для каждой скачанной по ссылке страницы. Эти вектора включают абсолютные частоты встречаемости терминов (слов) в тексте запроса или HTML-страницы. При этом в вектора не включаются частоты стоп-слов, т.е. слов, не несущих смысловой нагрузки, например, предлогов, общеупотребимых слов и т.п. (для этого используется специальный словарь стоп-слов). Для учета встречаемости терминов в разных морфологических формах при построении векторов используются основы терминов, выделенные с помощью стемминга (отсечения окончаний и суффиксов слов по определенным правилам).

Релевантность страницы запросу вычисляется как значение косинусной меры между векторами запроса $\vec{q}$ и страницы (документа) $\vec{d}$ по формуле 1.

$$\cos(\theta) = \frac{\vec{q} \cdot \vec{d}}{\|\vec{q}\| \times \|\vec{d}\|} = \frac{\sum_{i=1}^{n} q_i \times d_i}{\sqrt{\sum_{i=1}^{n} q_i^{\,2}} \times \sqrt{\sum_{i=1}^{n} d_i^{\,2}}} \quad (1),$$

где n – длина вектора (число учитываемых терминов), а i – позиция термина в векторе.

5 Извлечение информации

Для заполнения контента ИНИР собирается информация из таких источников, как сайты организаций, ассоциаций, проектов и конференций, порталы знаний, социальные научные сети 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 класса *Информационный ресурс*.

Для каждого из этих классов создается свой метод извлечения информации, включающий набор шаблонов, генерируемых на основе онтологии. Для повышения полноты извлечения информации вариативность этих шаблонов увеличивается за счет использования альтернативных терминов из тезауруса (синонимов и гипонимов), а также слов и словосочетаний, предлагаемых экспертом.

Модуль извлечения информации осуществляет анализ интернет-ресурсов, скачанных по ссылкам, заданным в БД СИР. Документы в сети Интернет могут быть представлены в различных форматах (HTML, DOC, PDF, TXT и др.). Но так как HTML является основным форматом для представления информации в Интернет, предлагаемые методы извлечения информации ориентированы на работу с HTML-страницами.

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 [8]. На основе соответствующего шаблона выполняется анализ DOM-дерева каждой страницы и извлечение описанной этим шаблоном информации.

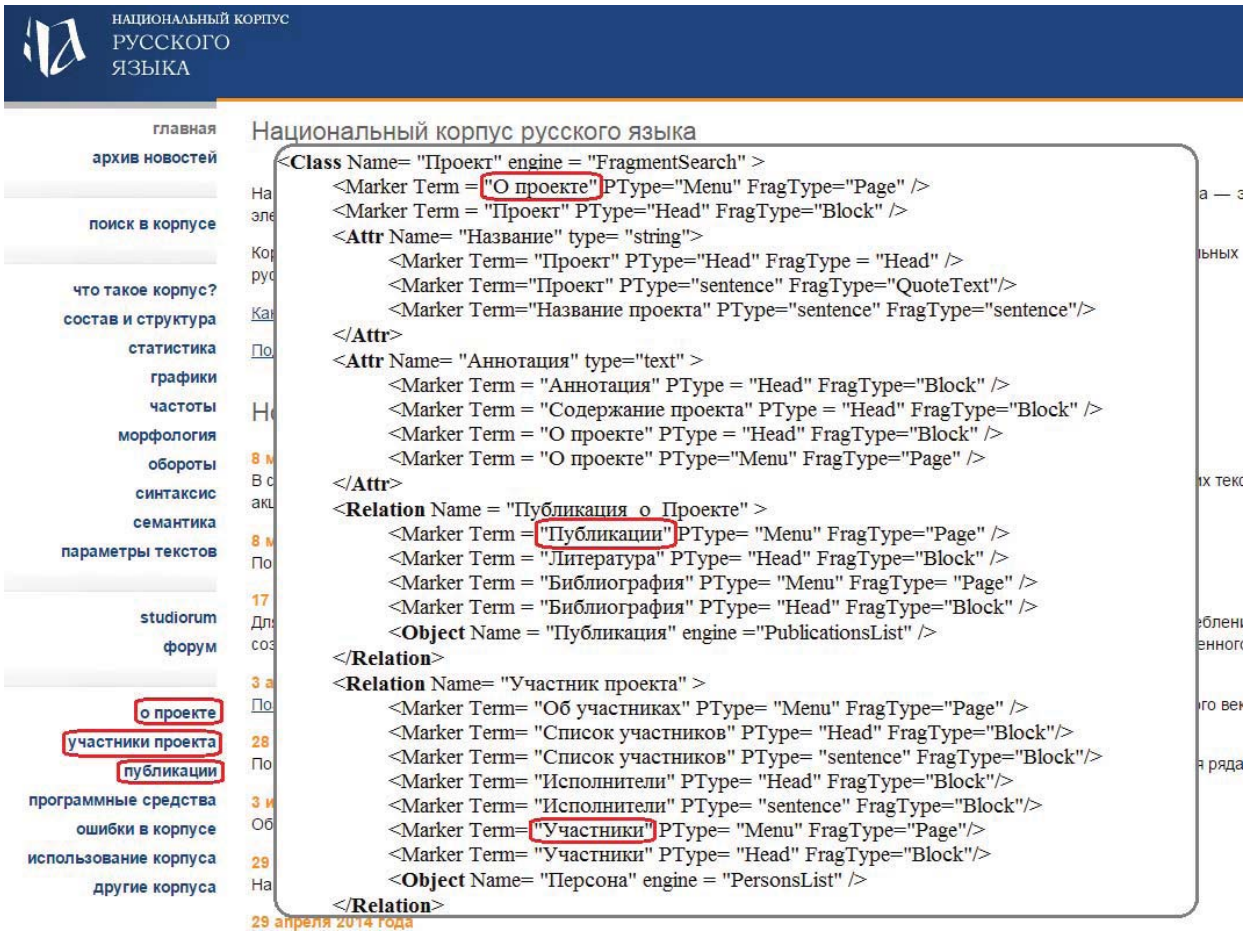

Рис.2. Извлечение информации с сайта с использованием шаблона

Шаблон представляет собой XML-документ, в котором для объектов, отношений и атрибутов онтологии указаны маркеры, сигнализирующие о расположении данного объекта, отношения или атрибута на HTML-странице. В шаблонах для каждого типа извлекаемой информации указываются обработчики, реализующие алгоритмы обхода и анализа соответствующих фрагментов интернет-страниц.

Важно отметить, что информация о сущностях, представляющих интерес для пользователей ИНИР, может быть задана различными способами. Например, информация о проекте, может быть представлена сайтом проекта, разделом сайта организации или персоны или публикацией, описывающей проект. Для каждого из этих способов представления на основе класса онтологии *Проект* строится отдельный шаблон.

На Рис. 2 представлен фрагмент шаблона для извлечения информации о проекте с сайта научно-исследовательского проекта. Этот шаблон позволяет извлекать такие атрибуты класса *Проект*, как «Название» и «Аннотация», а также аргументы отношений «Публикация о проекте» и «Участник проекта», т.е. объекты, описывающие соответственно публикации о проекте и участников проекта. Каждый шаблон, предназначенный для извлечения информации, 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.

Параметр маркера **Term** позволяет задать термин тезауруса, характеризующий смысл извлекаемой информации (например, список публикаций или участники проекта); параметр **PType** задает тип фрагмента, в тексте которого должен располагаться указанный термин (например, меню или заголовок), а параметр **FragType** – тип фрагмента, из которого будет извлекаться информация (например, блок или страница).

Следует заметить, что для каждого языка, используемого в ИНИР, задается свой набор маркеров. Для формирования таких маркеров используются соответствующие эквивалентные термины на разных языках, предоставляемые многоязычным тезаурусом, а также термины предложенные экспертом. Использование таких многоязычных шаблонов позволит извлекать информацию не только из интернет-ресурсов, в которых вся информация представлена на одном языке, но и из интернет-ресурсов, содержащих информацию на различных языках.

Как было сказано выше, шаблоны содержат также названия обработчиков, которые будут извлекать информацию из фрагментов, выделенных с помощью маркеров. Названия обработчиков указываются в соответствующих блоках шаблона (**Attr** и **Object**) после параметра engine.

Так как на различных сайтах объекты одного и того же класса могут представляться по-разному, в шаблоне для описания таких объектов могут использоваться различные маркеры. Например, на сайте организации позиционирование текста с информацией об объектах класса *Проект* может задаваться следующими маркерами: *проекты*, *гранты*, *прикладные разработки*, *практические разработки*, *основные направления*, *направления исследований*, *программы и проекты*; на сайте персоны – *участие в научных проектах*; на сайте проекта – *проект*, *о проекте*.

Поясним теперь, как происходит процесс извлечения информации из HTML-страницы.

Значениями параметров **PType** и **FragType** маркера могут быть различные элементы HTML-страницы, такие как меню, заголовок, блок. Для того чтобы обрабатывать маркеры такого типа, на анализируемой HTML-странице необходимо уметь выделять ее основные элементы. Как было сказано выше, страница представляется в виде DOM-дерева, в котором можно выделить меню и основной контент страницы. (Под основным контентом понимается, собственно, наполнение страницы, в частности, тексты, размещенные на ней.) Оставшиеся элементы (к ним относятся реклама, комментарии, «шапка страницы», «низ страницы» и т.п.) не представляют интереса для задачи извлечения информации. Далее в основном контенте выделяются заголовки и семантические блоки.

Для извлечения основных элементов используются эвристики. Так, эвристика, извлекающая меню, осуществляет поиск поддерева в DOM-дереве страницы, которое по своей структуре похоже на список ссылок, ведущих на текущий сайт. Другие эвристики служат для выделения основного контента на странице. Они основаны на идее, что основной контент содержит небольшое количество тегов разметки и большой объем текста относительно объема всей страницы.

После анализа HTML-страницы и выделения основных элементов выполняется поиск подходящего шаблона на основе маркеров блока **Class** и извлечение информации в соответствии с этим шаблоном. Для каждого блока в шаблоне (**Class**, **Attr**, **Relation**, **Object**) с помощью описывающих их маркеров извлекаются соответствующие сегменты, тип которых задается параметром **FragType**. Таким сегментом может быть, как элемент текущей страницы, так и другая HTML-страница, которая в процессе работы загружается и также анализируется.

Далее, в зависимости от характеристик извлекаемого блока и параметров шаблона, возможны следующие ситуации:

- Если у блока в шаблоне есть дочерние блоки, они обрабатываются рекурсивно. При этом область, в которой будет производиться поиск соответствий для маркеров, сужается до текущего сегмента.
- Если для блока в шаблоне указан специальный обработчик, тогда последнему на обработку передается текущий сегмент, приведенный к текстовому виду. После этого извлеченная обработчиком информация преобразуется к требуемому формату.
- Если блок в шаблоне является листовым, тогда формируется значение соответствующего объекта или атрибута из текущего сегмента на странице.

Таким образом, в процессе рекурсивной обработки шаблона с помощью соответствующих маркеров выделяются сегменты, которые затем обрабатываются специальными обработчиками, указанными в шаблоне. При этом формируется объект заданного онтологического класса и его связи с другими объектами.

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. Шаблон, построенный на основе класса *Проект* (см. Рис.2), 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.

На Рис.2 показано использование двух таких обработчиков PublicationList и PersonList. Первый из них предназначен для разбора списка публикаций, а второй – для обработки информации о персонах.

6 Занесение информации в контент ИНИР

Как было сказано выше, извлекаемая из интернет-ресурсов информация представляется в виде семантической сети информационных объектов, т.е. ориентированного мультиграфа. Интеграцию полученного графа в ИНИР выполняет модуль занесения информации.

Задача автоматического занесения информации довольно нетривиальна, так как кроме целостности данных необходимо обеспечить также согласованность и связанность вводимой информации с информацией, уже имеющейся в контенте ИНИР. Объекты научной деятельности, в соответствии с онтологией, характеризуются не только своими атрибутами, но и графом связанных объектов. Таким образом, задача занесения информации заключается в корректном объединении существующей в ИНИР семантической сети с графом, полученным при извлечении информации из некоторого ресурса. Для этого каждый вносимый информационный объект нужно проверить на корректность и существование в контенте ИНИР.

Проверка существования информационного объекта носит название идентификации. Суть ее состоит в разрешении неоднозначности, возникающей, когда для объекта из входного графа на основании значений его атрибутов невозможно однозначно сказать, будет ли он новой вершиной в семантической сети ИНИР или дополнит информацию об одной из уже имеющихся вершин.

Пусть $G=<V,E>$ — граф контента ИНИР, $g=<v,e>$ — граф объектов, извлеченных из некоторого интернет-ресурса. Вначале необходимо найти все общие вершины графов без учета ребер, т.е. только на основании значений атрибутов объектов. Пусть после этого граф g разделится на подграфы $g_1=<v_1,e_1>$ и $g_2=<v_2,e_2>$, где первый включает вершины, общие с графом G, а второй — вершины, при проверке которых возникла неоднозначность. Также на этом этапе можно выделить отдельное множество объектов-вершин из g, не входящее ни в g_1, ни в g_2, которые идентифицированы как новые, ранее не встречавшиеся в G. Объект может быть идентифицирован как новый только в том случае, когда у него полностью определен набор свойств и связей, обязательных для объектов заданного класса, и он отличается от таких же наборов свойств и связей всех других объектов того же класса, хранящихся в контенте ИНИР.

В дальнейшей проверке основную роль играют связи между вершинами подграфов g_1 и g_2. Необходимо одну за другой проверить все объекты-вершины v_2^i ($i=1,2,\ldots|v_2|$) подграфа g_2. Пусть объект v_2^i связан с вершинами подграфа g_1 ребрами-отношениями $e^i=\{e_1^i,e_2^i,\ldots,e_n^i\}$, а в силу имеющейся неоднозначности объекту v_2^i по набору значений его атрибутов можно сопоставить множество объектов $V^i=\{V_1^i,V_2^i,\ldots,V_k^i\}$ из графа G. Каждый из объектов V_j^i, в свою очередь, также связан некоторым (возможно пустым) множеством E_j^i ребер с вершинами графа g_1. Поочередно отбрасывая из множества V^i объекты V_j^i, имеющие менее l связей, однотипных со связями из e^i (т.е. являющихся экземплярами одного и того же онтологического отношения и связывающего V_j^i с теми же объектами из g_1), при $l=1,2,\ldots,n$ мы придем к одному из трех результатов: (1) множество V^i пусто, (2) множество V^i содержит более одного элемента или (3) множество V^i содержит ровно один элемент V_0^i. Последний случай означает, что v_2^i и V_0^i — общие вершины графов g и G, и объект v_2^i, таким образом, идентифицирован в контенте ИНИР как V_0^i. Вершина v_2^i после этого переносится в подграф g_1. Обходы подграфа g_2 повторяются итеративно и завершаются либо в случае, когда на предыдущей итерации не удалось идентифицировать ни одного объекта, либо если множество v_2 стало пустым. После завершения идентификации объектов проводится слияние графов по общим вершинам и добавление новых элементов в граф контента. В результате объединения могут появиться новые вершины в графе контента, новые ребра между уже существовавшими вершинами и/или новые атрибуты у таких вершин.

Заметим, что кроме информации о проектах, организациях, персонах, конференциях и публикациях, необходимо также заносить информацию о ее источниках, т.е. интернет-ресурсах (сайтах, порталах и т.п.), которая должна представляться в контенте ИНИР в виде объектов и отношений класса *Информационный ресурс*.

7 Заключение

Тематические интеллектуальные научные интернет-ресурсы позволяют исследователям значительно сократить время, требуемое для обеспечения доступа к информации по интересующим их тематикам и ее анализа. При этом эффективность использования каждого конкретного ИНИР напрямую зависит от полноты и корректности представленной в нем информации. Добиться такой полноты можно только за счет автоматизации процесса сбора информации. Для этих целей разрабатывается подсистема сбора информации из сети Интернет.

На данный момент реализованы все основные компоненты данной подсистемы и разработаны методы извлечения информации о проектах, персонах, организациях и событиях, включая сопутствующие шаблоны и обработчики,

реализующие извлечение информации о публикациях.

Литература

[1] Chen J., Chen H. A Structured Information Extraction Algorithm for Scientific Papers based on Feature Rules Learning // Journal of Software, Vol. 8, No. 1, January 2013. P. 55–62.

[2] 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.

[3] 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.

[4] 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.

[5] Hillmann D. Using Dublin Core, 2005. http://dublincore.org/documents/usageguide/

[6] Labský M., Svátek V., Nekvasil M., Rak D. The Ex Project: Web Information Extraction Using Extraction Ontologies // Knowledge Discovery Enhanced with Semantic and Social Information. Studies in Computational Intelligence. Berlin: Springer-Verlag, 2009. vol. 220, p. 71–88.

[7] Saggion H., Funk A., Maynard D., Bontcheva K. Ontology-based Information Extraction for Business Intelligence // Proceedings of the 6th international The semantic web (ISWC'07) and 2nd Asian conference on Asian semantic web conference (ASWC'07). Berlin, Heidelberg: Springer-Verlag, 2007. P. 843-856.

[8] Stenback J., Le Hégaret P., Le Hors A. Document Object Model (DOM) Level 2 HTML Specification // W3C Recommendation, 2003. http:// www.w3.org/TR/2003/REC-DOM-Level-2-HTML-20030109/

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

An Automatization of Collection of Information about Scientific Activity for Thematic Intelligent Scientific Internet Resources

Yury A. Zagorulko, Irina R. Akhmadeeva, Alexey S. Sery

The paper considers the problems of information collection and extraction for thematic intelligent scientific internet resources providing the systematization and integration of scientific knowledge, information resources and methods of intelligent information processing related to certain area of knowledge, as well as the content-based access to them. The approach to automatization of collection of information about scientific activity in the given knowledge area combining metasearch and knowledge extraction methods based on ontology and thesaurus is proposed. In accordance of this approach for every type of entities (ontology class) the specific methods of information collection and extraction adjustable to knowledge area and types of information resources is developed.

Each of these methods includes a set of patterns. In these patterns, for every kind of extracted information, markers defining its position are given as well as the engines implementing the algorithm of the analysis of the corresponding fragments of Web pages and extraction of the required information from them. These patterns are generated on the basis of the ontology. To improve the recall of information extraction, the patterns use alternative terms in different languages from thesaurus (synonyms and hyponyms) to describe the markers.