# ПОДХОД К АВТОМАТИЗАЦИИ СБОРА ОНТОЛОГИЧЕСКОЙ ИНФОРМАЦИИ ДЛЯ ИНТЕРНЕТ-ПОРТАЛА ЗНАНИЙ[1]

***О.И. Боровикова***

***olesya@iis.nsk.su***

***Ю.А. Загорулько***

***zagor@iis.nsk.su***

***Е.А. Сидорова***

***lena@iis.nsk.su***

***Институт систем информатики СО РАН,***

***Российский НИИ Искусственного Интеллекта, Новосибирск***

Обсуждается подход к автоматизации сбора онтологической информации о ресурсах, относящихся к предметной области Интернет-портала знаний. Специальная подсистема выполняет поиск релевантных документов, их содержательный анализ, индексирование и классификацию на основе предметного словаря и онтологии.

## *Введение*

Для решения задачи сведения ресурсов, относящихся к одной области знаний в единое информационное пространство, обеспечения возможности открытого и удобного доступа к ним, а также поддержки их целостности нами была предложена концепция специализированных Интернет-порталов знаний [1].

Основу портала знаний составляет онтология и соотнесенное с ней описание соответствующих сетевых ресурсов.

Особенность предложенной концепции состоит в том, что портал знаний обеспечивает доступ не только к собственным информационным ресурсам, но и поддерживает навигацию по заранее размеченным (проиндексированным) ресурсам, размещенным в сети Интернет. При этом информация о ресурсах накапливается коллекционером онтологической информации, т.е. специальной подсистемой портала знаний, осуществляющей сбор, анализ, оценку релевантности Интернет-ресурсов, а также их автоматическое индексирование и классификацию.

Коллекционер онтологической информации о ресурсах фактически выполняет функцию извлечения знаний и данных из сети Интернет [2].

В этом докладе подход к сбору онтологической информации рассматривается на примере портала знаний, служащего для поддержки научных исследований.

## *Онтология портала*

Для достаточно полного и целостного представления пользователя о выбранной отрасли знаний портал знаний включает следующие относительно независимые онтологии: онтологию науки и онтологию предметной области (ПО), описывающую конкретную отрасль знаний. Такое структурирование системы знаний на предметно-зависимые и предметно-независимые онтологии, значительно упрощает настройку портала на выбранную область научных знаний.

На Рис. 1. представлена система знаний, используемых в предлагаемом подходе.


*Рис. 1. Система знаний портала*

**Онтология науки** включает описывающие науку и научную деятельность классы понятий с заданными на них семантическими отношениями. Онтология науки условно разделена на онтологию

научной деятельности и онтологию научного знания.

**Онтология научной деятельности** включает общие классы понятий, относящиеся к организации научной деятельности: *Ученые, Организации, События, Публикации, Информационный ресурс*.

**Онтология научного знания** содержит метапонятия, задающие структуры для описания рассматриваемой предметной области, такие как *Раздел науки, Метод исследования и Объект исследования, Научный результат*.

**Онтология предметной области** отражает общие знания о ПО, такие как иерархия классов понятий, семантические отношения на этих классах.

**Предметные знания** – выделенная часть знаний о ПО, включающая только частные понятия и конкретные отношения.

**Онтология языка документов (словарь)** – это система языковых средств выражения онтологии ПО. Лингвистическая информация представлена в словаре с помощью функциональных групп лексических единиц, выделенных классов понятий и набора дополнительных атрибутов, отражающих специфику выражений: синонимы, омонимы, составные понятия и т.п.

**Языковые ресурсы** – это исходные данные для системы знаний, характеризующие предметную область пользователя. Обеспечить автоматическое извлечение знаний из этих данных является главной задачей эксперта при наполнении и настройке системы знаний портала.

Использование в качестве основы портала набора онтологий делает систему знаний портала легко расширяемой и настраиваемой – в нее могут интегрироваться как новые знания, так и новые типы информационных ресурсов.

### *Коллекционер онтологической информации*

Под сбором онтологической информации о ресурсах подразумевается как поиск ссылок на новые релевантные предметной области портала документы, так и фиксирование информации об этих документах как об экземплярах понятия онтологии *информационный ресурс*. Последнее состоит в определении значений атрибутов ресурса (название, ссылка, язык, тип доступа и т.д.) и задании связей с другими понятиями онтологии портала (организациями, публикациями, событиями и т.д.).

Коллекционер онтологической информации о ресурсах (Рис.2) включает два основных модуля: модуль сбора информации и модуль индексирования и классификации.

Модуль сбора информации осуществляет поиск Интернет-документов по ссылкам, заданным в специальной базе данных, и определяет их релевантность тематике портала.

Модуль индексирования и классификации, используя онтологию и предметный словарь, строит содержательный индекс для каждого документа и определяет раздел науки, к которому он относится.

### *Модуль сбора информации*

Модуль сбора информации включает следующие компоненты:

- базу данных ссылок на документы;
- словарь терминов (ключевых слов);
- поискового робота.

Поисковый робот обеспечивают поиск Интернет-документов (полуструктурированных и неструктурированных ресурсов) по ключевым словам на сайтах и страницах, ссылки на которые заданы в специальной базе данных (см. Рис.2).

База данных ссылок может пополняться как вручную (настройщиком-экспертом портала), так и автоматически (за счет ссылок, обнаруженных в документах). Кроме того, эта база данных может пополняться поисковым механизмом портала, который запускается с определенной периодичностью с целью обнаружения ссылок на новые ресурсы (сайты или порталы), релевантные тематике портала. Обеспечивается также возможность ввода параметров устаревания ссылки и периодичности повторной закачки документов по этой ссылке.

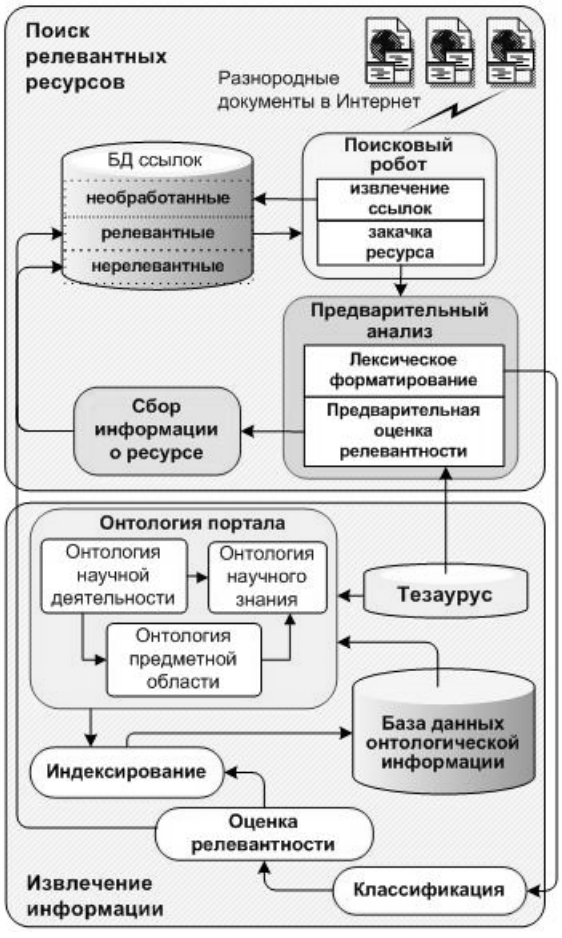


*Рис. 2. Схема сбора онтологической информации о ресурсах*

В основе поиска новых документов по заданным ссылкам лежит идея последовательного отсева документов согласно указанным при настройке портала критериям релевантности. При этом формируется поисковый образ документа, в котором с помощью предметного словаря (тезауруса) задается набор терминов, относящихся к предметной области и/или онтологии портала, которые должны содержаться в релевантном документе. Кроме этого поисковый образ может включать описание свойств документа: дату создания (редактирования), язык, тип ресурса и т.п.

Релевантность документа зависит от таких его параметров как:

1) расположение ключевых слов в html-тэгах документа;
2) расположение ключевых слов в выделенных фрагментах текста (заголовок, аннотация и т.п.);
3) встречаемость ключевых слов в адресе ссылки или домена;
4) вес ключевых слов в текстовом содержимом документа.

Работа модуля сбора информации разбивается на три этапа: анализ релевантности найденного по ссылке документа, поиск в документе ссылок на другие документы и сбор информации о документе.

На первом этапе с учетом параметров 1-3 определяется принадлежность документа поисковому образу согласно предварительному условию релевантности: “наличие хотя бы одного ключевого слова поискового образа в текстовом содержимом (html-коде) Интернет–документа”. При этом учитывается также и положение ключевого слова в документе. Для этого каждому выделенному фрагменту документа (заголовок страницы, заголовки текста на странице, список ключевых слов страницы, имя гиперссылки, название изображения и др.) приписывается вес, означающий степень важности встречаемости ключевого слова в данном месте документа.

Окончательное решение о релевантности и ее числовой оценке принимается после анализа его полного текста согласно критерию 4. Для этого текстовые ресурсы полностью скачиваются для определения статистики встречаемости ключевых слов в документе и оценки их релевантности на основе этой статистики.

Если полный текст не доступен, то решение о релевантности принимается по имеющейся аннотации. Решение о релевантности графических и мультимедиа-ресурсов принимается на основании всей имеющейся о них текстовой информации, например, подписей и аннотаций.

На втором этапе осуществляется анализ гиперссылок, обнаруженных в документе. Гиперссылки на документы, дополняющие информацию, размещенную в текущем документе, сохраняются в базе данных ссылок с целью их последующей обработки.

На третьем этапе осуществляется сбор информации о документе на основе представленной в нем метаинформации и его текстового содержимого. Собранная о документе (ресурсе) информация (в том числе, ссылка на него) сохраняется в базе данных онтологической информации (БД ОИ).

Дальнейший сбор информации продолжается на этапе индексирования, где происходит выделение из текста объектов и связей, описанных при помощи онтологии.

### *Модуль индексирования и классификации*

Современные системы обработки и анализа текстов на естественном языке используют либо статистический, либо лингвистический подход [3]. Специфика нашей задачи требует использования обоих подходов. В связи с этим модуль индексировании и классификации включает следующие компоненты:

- модуль лексического форматирования;
- словарь значимой лексики;
- набор обработчиков, отвечающих за автоматизированное наполнение и обучения словаря;
- модуль классификации;
- модуль индексирования документов.

На вход модуля индексирования и классификации поступает текст ресурса (как правило, в html-формате). Модуль лексического форматирования преобразует этот текст в «плоский», исключая из него служебную информации, требуемую для представления ресурса в Интернет. Мы будем считать, что текст не содержит аграммотичностей (т.е. будем просто игнорировать в тексте неизвестные лексемы).

Результатом работы модуля будет семантический индекс документа, т.е. набор объектов и отношений, представляющих его содержание в терминах онтологии портала. Индекс документа заносится в базу данных онтологической информации.

#### *Словарь*

Создание словаря является одним из самых трудоемких процессов при применении лингвистических методов анализа текстов на естественном языке. Специфика поставленной задачи определила требования, предъявляемые к словарю:

- Словарь должен содержать морфологическую информацию о терминах. Это требование с одной стороны связано с проблемой повышения качества оценки релевантности текстовых ресурсов, с другой – с необходимостью увеличить точность семантического анализа.
- Словарь должен хранить статистическую информацию. Так как при создании портала знаний, как правило, изначально имеется

большая выборка ресурсов, размеченная или соотнесенная разделам науки, то, используя классические методы обучения можно сразу получить начальное наполнение словаря, которое в противном случае пришлось бы вводить вручную многочисленным специалистам. Помимо этого, такая информация позволит использовать статистические методы классификации (рубрикации) для определения общей тематики ресурса (т.е. к какому разделу науки относится данный ресурс).

- Словарь должен хранить семантическую информацию, которая позволит связать элементы словаря с онтологическими классами проблемной и предметной области и которая в дальнейшем, должна будет использоваться на стадии семантического анализа.

Для этих целей используется технологический комплекс Алекс+, предназначенный для создания предметно-ориентированных словарей. Комплекс позволяет включать в словари как статистическую, так и семантическую информацию и поддерживает технологию автоматического наполнения словаря на основе обучающей выборки.

Словарная подсистема обеспечивает:

- морфологический анализ текста;
- сборку словокомплексов на основе системы правил-шаблонов;
- просмотр конкорданса;
- создание и редактирование иерархии тем (разделов науки);
- обучение словаря, т.е. автоматическое наполнение словаря терминами и словокомплексами на основе обучающего корпуса текстов;
- выявление стоп-терминов;
- классификацию текстов на основе ведущейся статистики.

### *Классификация*

Наличие словарных статистических показателей делает возможным применение классических методов классификации – процесса распознавания темы (набора тем) текста. Модуль классификации осуществляет классификацию по разделам науки.

На данный момент используется следующая процедура классификации: для всех значимых терминов текста для каждой темы (раздела науки) вычисляются суммы весов тех терминов, веса которых превышают шумовой уровень лексики и вычитается сумма обратных весов тех терминов, веса которых ниже шумового уровня лексики. Т.о. при анализе учитывается не только «положительная», но и «отрицательная» информация о соответствии термина теме.

В дальнейшем можно будет от простых функций распознавания, переходить к более сложным, учитывающим корреляцию пар лексем, выявление сложной значимой лексики и конструкций.

Текст относится к теме/темам, получившим значение функции выше некоторого порога. Значение этого порога определяет релевантность ресурса.

Таким образом, в результате работы модуля классификации определяется не только набор разделов науки, к которым относится текст, но и степень релевантности данного документа выявленным разделам, что дает основание дать команду на продолжение анализа текста (переход к индексированию) или же о прекращении анализа и исключения данного ресурса из списка релевантных.

### *Индексирование*

Под *индексированием* понимается процесс извлечения из текста документа объектов и связей, соответствующих понятиям и отношениям онтологии. Выделение таких объектов и связей осуществляется на этапе семантического анализа текста.

Онтология портала задает иерархию классов (здесь под классами понимаются не только понятия онтологии, но и отношения) и каждому классу в словаре сопоставляется группа терминов (причем термин одновременно может быть соотнесен нескольким классам).

На вход модуля индексирования поступает множество ключевых понятий, выделенных словарным компонентом системы при лексическом анализе текста. Дальнейший алгоритм автоматического индексирования документов реализуется в три этапа.

На этапе *сегментации* осуществляется жанровая декомпозиция текста, которая определяет тематические разделы, ограничивая возможную смысловую нагрузку той или иной части текста документа. Этот этап тесно связан с этапом лексического форматирования текста, где, используя знания о специальных символах разметки документов (такие как тэги), можно определить значимость того или иного фрагмента текста.

Последующая обработка документа представляет собой процесс извлечения релевантной информации на основе ключевых понятий.

*Идентификация объектов*. На этом этапе определяются все возможные атрибуты понятия, позволяющие уточнить объект, описываемый данным понятием. Кроме того, делается попытка сопоставить найденное понятие объектам, хранящимся в БД ОИ и полученным при анализе ранее поступивших ресурсов.

Непосредственно на этапе *семантического анализа* осуществляется связывание объектов на основе семантической сочетаемости соответствующих им понятий, а также с учетом проективности (связи не должны пересекаться) и

связности (при возникновении многовариантности выбираются разбиения, содержащие минимальное количество несвязанных элементов) фрагментов текста, покрываемых данными понятиями.

Следует отметить, что здесь не проводится глубокий семантический анализ, т.к. связывание осуществляется только для тех пар объектов, для которых в онтологии представлены соответствующие связи.

Индекс документа помещается в БД ОИ; при этом, если включенные в индекс объекты уже существуют в БД, то значения некоторых их атрибутов могут уточняться. Возникающие противоречия полученных при индексировании результатов с уже существующим в БД разрешаются администратором портала или экспертами.

Подобный описанному в докладе механизм индексирования текстов был успешно опробован в системе интеллектуализации документооборота InDoc [4].

### *Заключение*

В докладе предложен подход к автоматизации сбора и накопления онтологической информации о ресурсах, релевантных предметной области Интернет-портала знаний.

Ближайшими целями авторов является апробация предложенного подхода. В частности, в настоящее время ведется реализация коллекционера онтологической информации для портала знаний, обеспечивающего содержательный доступ широкому кругу пользователей к информационным ресурсам по археологии и этнографии [5]. Кроме того, планируется применение данного подхода при разработке Интернет-портала, предназначенного для информационной и интеллектуальной поддержки инновационной деятельности в сибирском регионе.

### *Список литературы:*


1) Боровикова О.И., Загорулько Ю.А. Организация порталов знаний на основе онтологий. // Труды международного семинара Диалог'2002 “Компьютерная лингвистика и интеллектуальные технологии”. Протвино, 2002. Т.2, С.76-82.
2) Тихомиров И.А. Распознавание интерфейсов Интернет-ресурсов на основе использования неоднородных семантических сетей // Труды конференции КИИ-2004. М.: Физматлит, 2004. Т.1, С.179-185
3) Хорошевский В.Ф. Управление знаниями и обработка ЕЯ-текстов // Труды 9-й национальной конференции по искусственному интеллекту - КИИ'2004. Москва: Физматлит, 2004. Т.2, С.565-572.
4) Загорулько Ю.А., Кононенко И.С., Сидорова Е.А., Костов Ю.В. Подход к интеллектуализации документооборота // "Информационные технологии", 2004. № 11, С. 2-11.
5) Боровикова О.И., Булгаков С.В., Загорулько Ю.А., Сидорова Е.А., Холюшкин Ю.П. Концепция интеллектуального интернет-портала знаний для доступа к информационным ресурсам по археологии и этнографии // Труды VI-й международной конференции "Проблемы управления и моделирования в сложных системах". Самара: Самарский Научный Центр РАН, 2004. С. 215-220.



[1] Работа выполняется при финансовой поддержке РФФИ (проект № 04-01-00884а) и Президиума СО РАН (Междисциплинарный интеграционный проект № 149).