СПИСОК ЛИТЕРАТУРЫ

1. Front-end and back-end – Режим доступа: http://en.wikipedia.org/wiki/Front-end_and_back-end – Загл. с экрана.
2. FAQ appendix 1: как писать сервера – fido7.ru.unix.prog. Группы Google – Режим доступа: http://groups.google.ru/group/fido7.ru.unix.prog/browse_thread/thread/e8f8edf4f2f2447b/ – Загл. с экрана.
3. The Apache HTTP Server Project – Режим доступа: http://httpd.apache.org/ – Загл. с экрана.
4. Уэйнрайт П. Apache для профессионалов. – М.: Wrox Press Ltd, 2001. – 474 с.
5. Nginx – Режим доступа: http://sysoev.ru/nginx/ – Загл. с экрана.
6. PHP: Hypertext Preprocessor – Режим доступа: http://www.php.net/ – Загл. с экрана.
7. Веллинг Л., Томсон Л. Разработка Web-приложений с помощью PHP и MySQL. – М.: Вильямс, 2007. – 880 с.
8. The world's most popular open source database – Режим доступа: http://www.mysql.com/ – Загл. с экрана.
9. Дюбуа П. MySQL. – М.: Вильямс, 2007. – 1168 с.
10. Siege – Режим доступа: http://www.joedog.org/JoeDog/Siege – Загл. с экрана.
11. Smarty : Template Engine – Режим доступа: http://www.smarty.net/ – Загл. с экрана.
12. Caching Tutorial for Web Authors and Webmasters – Режим доступа: http://www.mnot.net/cache_docs/ – Загл. с экрана.
13. Обработка ошибки 404 – Режим доступа: http://lekx.ru/modules/myarticles/article.php?storyid=515 – Загл. с экрана.
14. Memcached: a distributed memory object caching system – Режим доступа: http://www.danga.com/memcached/ – Загл. с экрана.
15. FastCGI – Режим доступа: http://ru.wikipedia.org/wiki/FastCGI – Загл. с экрана.

Поступила 23.04.2008 г.

Ключевые слова:

WEB-приложение, быстродействие, кэширование, динамическая страница, статический файл, производительность.

УДК 002.53:004.89

АВТОМАТИЗАЦИЯ СБОРА ОНТОЛОГИЧЕСКОЙ ИНФОРМАЦИИ ОБ ИНТЕРНЕТ-РЕСУРСАХ ДЛЯ ПОРТАЛА НАУЧНЫХ ЗНАНИЙ

Ю.А. Загорулько

Институт систем информатики им. А.П. Ершова СО РАН, г. Новосибирск
E-mail: zagor@iis.nsk.su

Предлагается подход к автоматизации сбора онтологической информации об Интернет-ресурсах, релевантных предметной области портала научных знаний. Специальная подсистема выполняет поиск ресурсов (документов), оценку их релевантности, содержательный анализ, индексирование и классификацию с использованием предметного словаря и онтологии предметной области.

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 этой статье подход к автоматизации сбора онтологической информации об Интернет-ресурсах рассматривается на примере портала знаний, служащего для поддержки научных исследований.

Система знаний портала

Базис системы знаний портала (см. рис. 1) составляет онтология, которая не только обеспечивает формальное представление системы понятий предметной области (ПО) портала, но и интеграцию в его информационное пространство релевантных информационных ресурсов.

Формально онтология портала знаний может быть описана семеркой вида:

$$O = \langle C, A, R_C, T, D, R_A, F \rangle,$$

где C – множество классов, описывающих понятия некоторой предметной или проблемной области;

A – множество атрибутов, описывающих свойства понятий и отношений; $R_C=\{r_C|r_C\subseteq C\times C\}$ – множество отношений, заданных на классах (понятиях); T – множество стандартных типов значений атрибутов (string, integer, real, date); D – множество доменов (множеств значений стандартного типа string); $R_A=R_{AT}2R_{AD}$, где $R_{AT}\subseteq A\times T$ – отношение, связывающее атрибуты и типы данных, из которых они могут принимать свои значения, $R_{AD}\subseteq A\times D$ – отношение, определяющее для каждого атрибута его дискретное множество значений (домен); F – множество ограничений на значения атрибутов понятий и отношений.

С содержательной точки зрения онтология портала служит для представления понятий, необходимых как для описания определенной области знаний, так и выполняемой в рамках нее научной деятельности. В связи с этим онтология портала знаний включает следующие онтологии: онтологию научной деятельности, онтологию научного знания и онтологию предметной области (ПО), описывающую конкретную отрасль знаний.

Рис. 1. *Система знаний портала*

Онтология научной деятельности и онтология научного знания являются предметно-независимыми, базовыми онтологиями. Они включает описывающие науку и научную деятельность классы понятий с заданными на них семантическими отношениями.

Онтология научной деятельности включает общие классы понятий, относящиеся к организации научной деятельности: *Исследователь, Организация, Деятельность, Событие, Публикация*. В эту онтологию также включен класс *Информационный ресурс*, который служит для описания информационных ресурсов, представленных в сети Интернет.

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

Онтология предметной области описывает область научных знаний, определенную тематикой портала. Она включает четыре базовые иерархии понятий, построенные на метапонятиях онтологии научного знания: иерархию разделов науки, иерархию объектов исследования, иерархию методов исследования и иерархию научных результатов. Все эти иерархии связаны между собой посредством ассоциативных отношений, часть которых наследуется из базовых онтологий, а часть отражает специфику данной предметной области.

Вводя формальные описания понятий проблемной и предметной области в виде понятий и отношений между ними, онтология портала задает структуры для представления реальных объектов (экземпляров понятий) и связей между ними (экземпляров отношений), составляющих **информационное наполнение портала знаний**.

Таким образом, данные на портале представлены в виде множества разнотипных информационных объектов и связей между ними. Информационный объект (ИО) – это структурированная совокупность данных, представляющая описание некоторого объекта выбранной области знаний или релевантного ей информационного ресурса. Каждый ИО соответствует некоторому понятию онтологии (является экземпляром этого понятия) и имеет заданную им структуру. Между конкретными информационными объектами могут существовать связи, семантика которых определяется отношениями, заданными между соответствующими понятиями онтологии.

Информация о ресурсах хранится в виде данных, соотнесенных с понятиями онтологии. Каждый ресурс соответствует такому понятию онтологии научной деятельности, как Информационный ресурс, а контекстное описание конкретного ресурса (т. е. информация о ресурсе как таковом) хранится в БД и включает экземпляр данного понятия и набор экземпляров отношений, связывающих данное понятие с другими понятиями онтологии. К атрибутам информационного ресурса относятся: название, ссылка, язык, тип доступа и т. п.; ресурс может быть связан отношениями с организациями, учеными, публикациями, событиями, разделами науки и т. д.

Содержание ресурса представляется набором фактов – экземпляров отношений понятий онтологии, найденных в тексте документа. Способ выражения фактов, извлекаемых из текстов на различных языках, задается с помощью компонента системы знаний портала – онтология языка документов.

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.

Технология сбора онтологической информации

Технология сбора онтологической информации о ресурсах включает два основных этапа: поиск в Интернете новых релевантных предметной области портала ресурсов (документов) и фиксирование информации о них как об экземплярах понятия онтологии *Информационный ресурс* в базе данных портала. 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.).

На рис. 2 показана общая схема поиска Интернет-ресурсов и извлечения из них значимой информации. Согласно этой схеме подсистема сбора онтологической информации о ресурсах 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-ресурсов (полуструктурированных и неструктурированных документов) 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.

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:

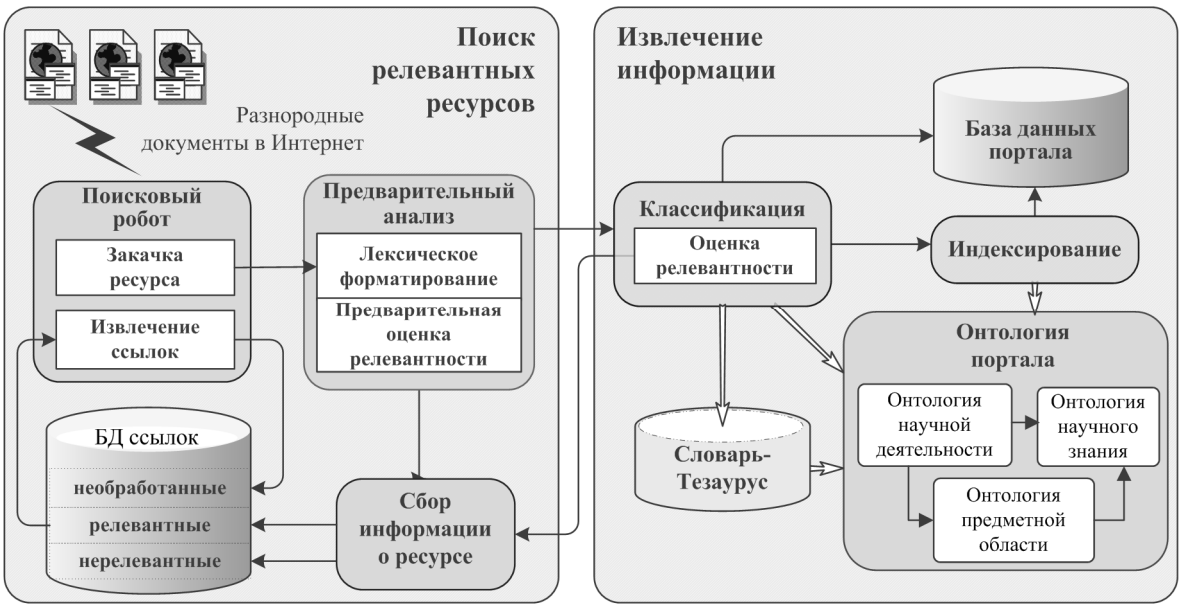

Рис. 2. *Схема сбора онтологической информации о ресурсах*

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.

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 информацию, требуемую для представления ресурса в Интернет.

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 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 портала и которая в дальнейшем может использоваться на стадии семантического анализа.

Для этих целей используется технологический комплекс, предназначенный для создания предметно-ориентированных словарей [4] и использования их в различных приложениях. Этот комплекс позволяет включать в словари как статистическую, так и семантическую информацию и поддерживает технологию автоматического наполнения словаря на основе обучающей выборки.

В целом комплекс 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 функция распознавания: 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. Текст относится к теме/темам, получившим значение функции распознания 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 [5]. 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 портала 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 портала; при этом, если включенные в индекс объекты уже существуют в БД, то значения некоторых их атрибутов могут уточняться. Противоречия, возникающие при внесении в БД результатов индексирования, разрешаются администратором портала или экспертами.

Заключение

Предложен подход к автоматизации сбора и накопления онтологической информации об Интернет-ресурсах, релевантных предметной области портала научных 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 научным знаниям и информационным ресурсам по компьютерной лингвистике [6].

Работа выполняется при финансовой поддержке РГНФ (проект № 07-04-12149).

СПИСОК ЛИТЕРАТУРЫ

1. Боровикова О.И., Загорулько Ю.А. Организация порталов знаний на основе онтологий // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. семинара «Диалог'2002». – Протвино, 2002. – Т. 2. – С. 76–82.
2. Тихомиров И.А. Распознавание интерфейсов Интернет-ресурсов на основе использования неоднородных семантических сетей // Труды 9-й национальной конф. по искусственному интеллекту «КИИ'2004». – М.: Физматлит, 2004. – Т. 1. – С. 179–185.
3. Хорошевский В.Ф. Управление знаниями и обработка ЕЯ-текстов // Труды 9-й национальной конф. по искусственному интеллекту «КИИ'2004». – М.: Физматлит, 2004. – Т. 2. – С. 565–572.
4. Сидорова Е.А. Технология разработки тематических словарей на основе сочетания лингвистических и статистических методов // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. «Диалог'2005». – М.: Наука, 2005. – С. 443–449.
5. Загорулько Ю.А., Кононенко И.С., Сидорова Е.А. Семантический подход к анализу документов на основе онтологии предметной области // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. «Диалог'2006». – М.: Изд-во РГГУ, 2006. – С. 468–473.
6. Боровикова О.И., Загорулько Ю.А., Загорулько Г.Б., Кононенко И.С. Подход к построению портала знаний по компьютерной лингвистике // Системный анализ и информационные технологии: Труды II Междунар. конф. – Обнинск, 2007. – Т. 1. – С. 126–129.

Поступила 05.03.2007 г.

Ключевые слова:

Онтологическая информация, Интернет-ресурсы, портал научных знаний, предметный словарь, онтология предметной области, поиск ресурсов, оценка релевантности, индексирование и классификация.

УДК 688.518:622.276

ОПТИМИЗАЦИЯ ПРОЦЕССА ЦИФРОВОГО 3D-МОДЕЛИРОВАНИЯ МЕСТОРОЖДЕНИЙ НЕФТИ И ГАЗА

А.А. Захарова, М.А. Иванов

Институт «Кибернетический центр» ТПУ
E-mail: alen@cc.tpu.edu.ru, IvanovMA@tpu.ru

Предложена схема оптимизации процесса цифрового 3D-моделирования месторождений нефти и газа. Это позволяет сократить сроки проектирования разработки нефтегазовых месторождений, повысить качество и надежность создаваемых моделей. Разработанные программные средства автоматизируют различные этапы процесса моделирования и применяются при решении практических задач. Данные средства были опробованы при проектировании разработки ряда месторождений Томской области, а также при подготовке специалистов в области проектирования разработки нефтегазовых месторождений.

Сегодня нефтегазодобыча является одной из наиболее наукоемких и высокотехнологичных областей производства. Поэтому в ней в полной мере востребованы современные информационные технологии (ИТ), при помощи которых создаются цифровые трехмерные модели месторождений нефти и газа с целью оценки запасов и состояния разработки, а также прогнозирования технологических показателей для выбора наиболее оптимальной стратегии выработки залежей углеводородного сырья. Стремительное развитие компьютерных технологий позволяет использовать высокопроизводительные вычислительные машины совместно с разработанными программными технологиями для сбора, хранения, расчета, представления и анализа различного рода данных, относящихся к процессу трехмерного моделирования месторождений. Совокупность современных вычислительных систем и специализированных программных комплексов (ПК) – это необходимый инструмент для любой нефтегазодобывающей компании. Поэтому применение и развитие ИТ при моделировании состояния разработки месторождений весьма актуально [1].

В процессе создания трехмерных моделей месторождений нефти и газа можно выделить три основных этапа:

1. Геологическое (Г) моделирование.
2. Гидродинамическое (ГД) моделирование.
3. Анализ результатов моделирования с целью принятия решений по управлению проектированием и разработкой месторождений нефти и газа.

На рис. 1, в качестве примера, приведена схема процесса моделирования с применением программного обеспечении (ПО) компании Schlumberger.

Как видно из рис. 1, основными инструментами представленной компании при построении цифровых трехмерных моделей являются программные комплексы «Petrel» и «Eclipse» [2]. В процессе моделирования специалисту приходится обрабатывать, систематизировать и хранить большое количество входной и создаваемой в процессе работы информации. При этом средств управления и автоматизации данным процессом не предусмотрено в предложенной схеме, что существенно повышает временные затраты на проектирование разработки месторождений нефти