OSTIS-2016
(Open Semantic Technologies for Intelligent Systems)

УДК 004.891

ЕДИНАЯ ОНТОЛОГИЧЕСКАЯ ПЛАТФОРМА ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Филиппов А.А., Мошкин В.С., Шалаев Д.О., Ярушкина Н.Г.

Ульяновский государственный технический университет,
г.Ульяновск, Российская Федерация

al.filippov@ulstu.ru

PostForVadim@yandex.ru

do.shalaev@ulstu.ru

jng@ulstu.ru

В данной работе предложена методика разработки, а также основные принципы построения и архитектура единой платформы интеллектуального анализа данных на основе онтологии предметной области. Помимо этого рассмотрен процесс построения предметной онтологии с помощью встроенного в платформу универсального редактора на примере онтологии «Справочник боцмана».
Ключевые слова: онтология, платформа, data mining, семантика.

Введение

В процессе деятельности любой современной крупной организации возникает необходимость в проведении интеллектуального анализа данных. Для проведения такого анализа от специалиста требуется глубокое знание предметной области, умение использовать различные инструменты интеллектуального анализа и знание современных IT-технологий для адаптации существующих решений под особенности конкретной предметной области.

Желание максимально автоматизировать процесс обработки данных с учетом специфики предметной области вызывает потребность в едином универсальном инструментарии интеллектуального анализа данных, решающего различные задачи, основанные на накопленных знаниях и опыте экспертов в этой области:

- построение экспертной системы, обеспечивающей вывод рекомендаций в ходе принятия управленческих решений;
- моделирование и прогнозирование показателей деятельности компании по данным предыдущих периодов отчетности;
- поиск необходимой документации в объемном архиве организации с учетом ее специфики и т. д.

Современное производство нуждается в создании адекватных методов и инструментальных средств повышения эффективности и качества выпускаемой продукции. Для этого необходимы соответствующие модели и методы, ориентированные на описание объектных структур рассматриваемой предметной области, позволяющие идентифицировать, анализировать (включая наглядное представление) и манипулировать всем многообразием объектов и отношений, имеющихся в предметной области.

Ряд научных проблем, стоящих перед современными организациями и требующих системного решения:

- необходимость разработки семантического базиса анализа содержимого электронного хранилища информации;
- отсутствие интегративных концептуальных моделей, использующих различные подходы хранения знаний о предметной области;
- необходимость универсализации процесса автоматизированной обработки хранимых знаний;
- необходимость одновременного использования разноаспектных описаний особенностей рассматриваемой предметной области;
- необходимость решения проблемы учета нечеткости в человеческих рассуждениях.

Решение данных проблем может основываться на применении интеллектуальных методов и алгоритмов анализа содержимого электронного хранилища предприятия с целью организации логического вывода рекомендаций для поддержки

принятия управленческих, экономических и технических решений с учетом модели конкретной предметной области.

В связи с этим, в настоящее время актуальной является задача обеспечения специалистов организаций различных профилей универсальным инструментарием, позволяющим решать поставленные задачи с учетом специфики конкретной предметной области.

1. Онтология – как база знаний современной компании

Современные производственные организации обладают значительными по объему электронными хранилищами различного рода информации, связанной с деятельностью данного предприятия. Фактически, такое электронное хранилище содержит в себе историю предприятия, опыт и знания большого количества высококвалифицированных специалистов.

В работах [Норенков, 2009] и [Титов, 2009] отмечается, что при увеличении объема электронного хранилища затрудняется анализ его содержимого по заранее заданным реквизитам, а от лиц, принимающих решения, требуются навыки в области семантической обработки большого объема информации, а также глубокие знания предметной области. В результате опыт и знания, зафиксированные в таких хранилищах, остаются невостребованными, что приводит к принятию неверных управленческих, экономических и технических решений.

Учет специфики предметной области приводит к необходимости формирования прикладной онтологии особой структуры, включающей в себя систему понятий предметной области, семантические отношения между ними и функции интерпретации. Ведущие исследователи в области онтологических систем отмечают актуальность исследований, основанных на онтологическом подходе [Добров и др., 2006] [Гаврилова и др., 2000]. В трудах данных исследователей отмечается важность использования онтологического инжиниринга в процессе интеллектуального анализа данных.

Наиболее универсальной и полной с точки зрения охвата специфики предметной области во всем разнообразии отношений между ее объектами является модель интеграции онтологического и продукционного подходов представления знаний, позволяющая в процессе логического вывода соответствующих рекомендаций опираться на данные, представленные в виде онтологии и наиболее полно описывающие рассматриваемую предметную область.

В широком смысле, онтологии – это модели, являющиеся формой представления знаний о предметных областях в виде семантических информационно-логических сетей взаимосвязанных объектов, где в качестве главных элементов выступают понятия предметной области с их свойствами и отношения между объектами. Онтологии выполняют интегрирующую функцию, обеспечивая общий семантический базис в процессах принятия решений, интеллектуального анализа данных и единую платформу для объединения разнообразных информационных систем.

В данной работе предлагается подход к построению единой платформы проведения интеллектуального анализа данных, включающей в себя объектно-ориентированные, а также нечеткие модели и программные инструментальные средства работы с ними, с использованием онтологического системного анализа.

2. Общее описание платформы

2.1. Цели и задачи исследования

Базовой целью данного исследования является разработка универсальной web-ориентированной платформы, позволяющей:

1. Решать широкий круг задач на основе интеллектуального анализа данных (индексация, классификация, кластеризация, информационный поиск).
2. Адаптировать платформу к конкретной предметной области с помощью методов онтологического анализа и инженерии знаний (разработка, хранение и использование модели предметной области в виде прикладной онтологии с возможностью импорта в различные форматы).
3. Осуществлять экспертную оценку и информационную поддержку в принятии решений (использование механизмов дискрипционной логики).

На настоящий момент было разработано ядро онтологической платформы, обеспечивающее автоматизированное формирование онтологии предметной области, визуализацию ее структуры, а также процесс информационного поиска знаний. Помимо этого, были поставлены следующие задачи:

1. Расширение функционала платформы для осуществления функций контроля прав доступа, организации хранения исходных и полученных данных.
2. Предоставление конечным пользователям возможности использования платформы без специальной подготовки, разработка интуитивно-понятного интерфейса.
3. Предоставление конечным пользователям средств адаптации платформы под особенности конкретной предметной области.
4. Предоставление разработчикам возможности расширения функционала платформы без значительных временных затрат на изучение архитектуры платформы с помощью механизма плагинов.

В настоящее время не существует полных аналогов разрабатываемой платформы. Косвенным аналогом можно считать веб-ориентированную систему ClowdFlows, позволяющую пользователю настраивать этапы проведения интеллектуального анализа данных, формировать модели данных и проводить эксперименты.

Однако для расширения функционала данной системы и адаптации под конкретную предметную область необходимо использовать веб-сервисы WSDL. Данный подход является достаточно сложным для использования неподготовленными пользователями, так как требует определенных знаний в IT-сфере.

2.2. Архитектура платформы

Архитектура платформы интеллектуального анализа представлена на рисунке 1.

Как видно из схемы, серверная часть платформы реализована на языке программирования Java с использованием фреймворка Spring, который предоставляет широкие возможности для облегчения разработки корпоративных приложений.

Интерфейс платформы реализован с применением фреймворка Google Web Toolkit (GWT), который позволяет создавать RIA (Rich Internet Application) Ajax-приложения на основе Java-кода путем его компиляции в JavaScript-код.

Использование GWT позволяет использовать типобезопасный язык Java при разработке, что снижает количество потенциальных ошибок и делает возможным отладку кода интерфейсной части платформы. При этом на клиенте данный код выполняется в виде легковесного JavaScript-кода, который поддерживается всеми современными браузерами.

Платформа выполняется в контейнере сервлетов Jetty. Данный контейнер имеет модульную архитектуру, что позволяет использовать только необходимый возможности [Wilkins, 2008], снижая нагрузку на сервер.

Также Jetty хорошо масштабируется для обслуживания многих соединений, со значительным временем простоя между запросами и позволяет обслуживать большее количество пользователей.

В основу платформы интеллектуального анализа данных положен принцип микросервисной архитектуры [Lewis et. Al, 2014].

Микросервисная архитектура – подход, при котором единое приложение является распределенным и состоит из небольших сервисов, которые тесно взаимодействуют между собой, при этом каждый такой сервис работает в собственном процессе. Для осуществления взаимодействия сервисов часто используются легковесные механизмы удаленного вызова процедур.

Каждый сервис, как правило, реализует конкретный бизнес-процесс и выполняется независимо в изолированном контейнере. Данный подход позволяет реализовывать сервисы с помощью различных языков программирования и использовать различные хранилища данных.

Такая изолированность сервисов позволяет:

1. Повысить общую отказоустойчивость системы за счет выполнения сервисов в разных адресных пространствах.
2. Повысить масштабируемость системы за счет запуска нескольких экземпляров сервисов и балансировки нагрузки между ними.
3. Предоставить возможность использования различных операционных систем, языков программирования, технологий хранения данных и т. д., что позволяет использовать наилучшее решение поставленной задачи.
4. Уменьшить время простоя системы при внесении изменений, исправлении ошибок и других сервисных задач.

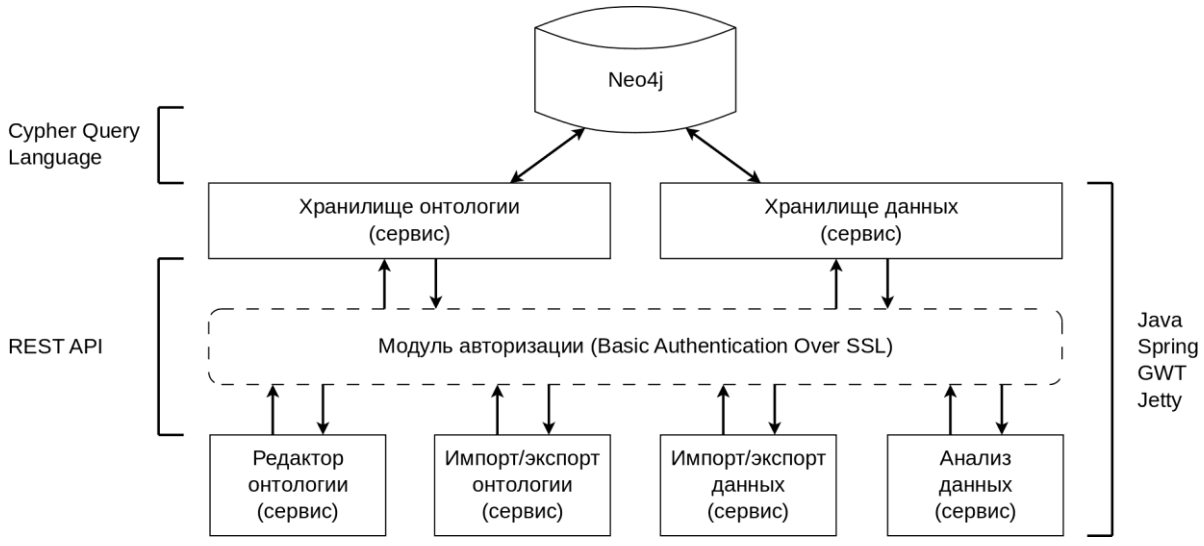

Рисунок 1 – Архитектура платформы интеллектуального анализа данных

Для взаимодействия сервисов в рамках рассматриваемой платформы используется механизм взаимодействия REST (Representational State Transfer – «передача репрезентативного состояния»), при котором вызов удаленной процедуры представляет собой обычный HTTP-запрос (GET, POST, PUT и т. д.), а необходимые данные пере-даются в качестве параметров запроса.

В качестве хранилища онтологий платформы интеллектуального анализа используется графовая база данных Neo4j. Данная СУБД обладает следующими преимуществами:

1. Нативный формат хранения графов.
2. Один экземпляр СУБД может обслуживать графы с миллиардами узлов и связей.
3. Может обрабатывать графы, которые полностью не помещаются в оперативной памяти.

Графо-ориентировнный язык запросов – Cypher.

Все перечисленные ресурсы, приложения и технологии являются бесплатными и свободно распространяемыми.

2.3. Процесс построения предметной онтологии

Рассмотрим процесс построения предметной онтологии с помощью встроенного в платформу универсального редактора на примере онтологии «Справочник боцмана».

Очень часто человек, плохо знакомый с предметной областью имеет информационную потребность, которую необходимо удовлетворить [Наместников и др, 2014].

В [Наместников и др, 2014] дано следующее определение информационной потребности пользователя: «Информационная потребность – это информационная неопределенность, которую пользователь хочет уменьшить посредством получения информации из системы информационного поиска». В нашем случае такой системой будет являться платформа интеллектуального анализа данных.

Все необходимые для удовлетворения информационной потребности пользователя данные содержатся в онтологии предметной области, которую, как правило, формируют эксперты данной предметной области.

Рассмотрим подробно данный процесс на примере онтологии «Справочник боцмана».

Корабль – сложный объект, состоящий из большого числа составных узлов и имеющий различные характеристики. У новичков в данной предметной области часто возникают вопросы, например, о том якорную цепь какого калибра, с какой разрывной нагрузкой и массой на 1 метр цепи установить на корабль определенного водоизмещения. Ответы на данные вопросы можно узнать, используя встроенный в редактор онтологии механизм информационного поиска знаний (рисунок 2), работка которого основана на содержимом онтологии предметной области (рисунки 3 и 4).

Рисунок 2 иллюстрирует пример запроса к системе для удовлетворения информационной потребности пользователя.

Из рисунка 3 видно, что встроенный в платформу интеллектуального анализа данных редактор онтологий позволяет создавать «схему» онтологии, которая состоит из классов и отношений между ними, а также заполнять данную онтологию знаниями, которые представлены индивидуалами (объекты классов) с определенными состояниями (объекты отношений) (рисунок 4).

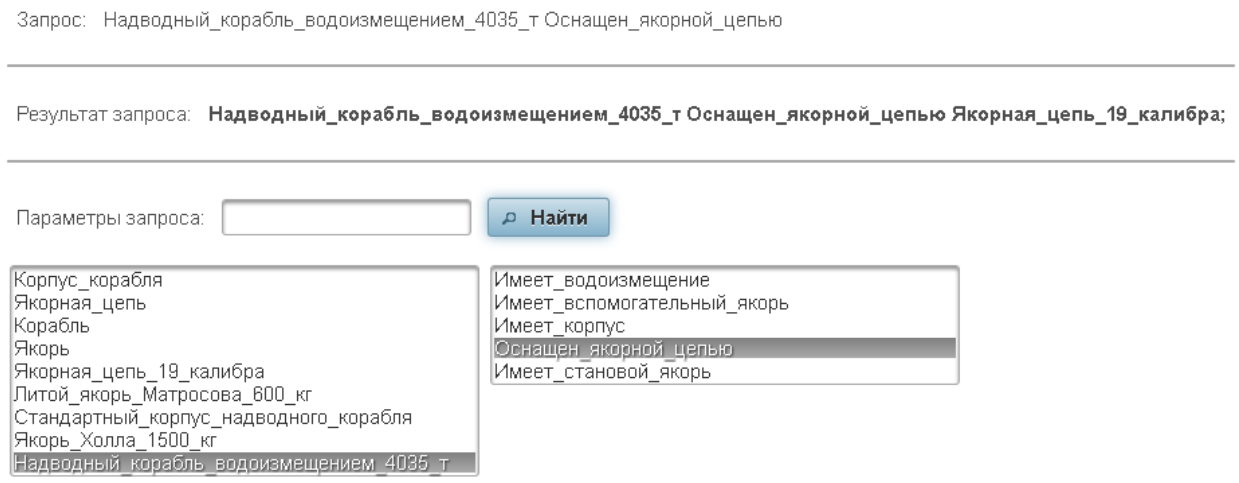

Рисунок 2 – Окно модуля информационного поиска. Пример запроса к системе.

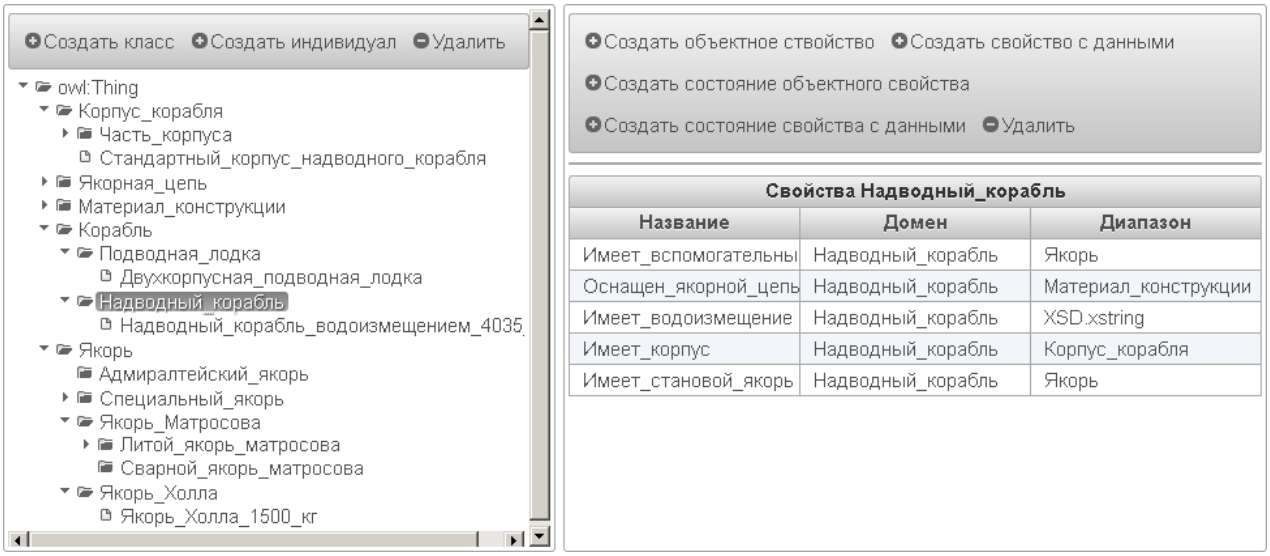

Рисунок 3 – Окно редактора онтологий. Класс «Надводный корабль» и его свойства

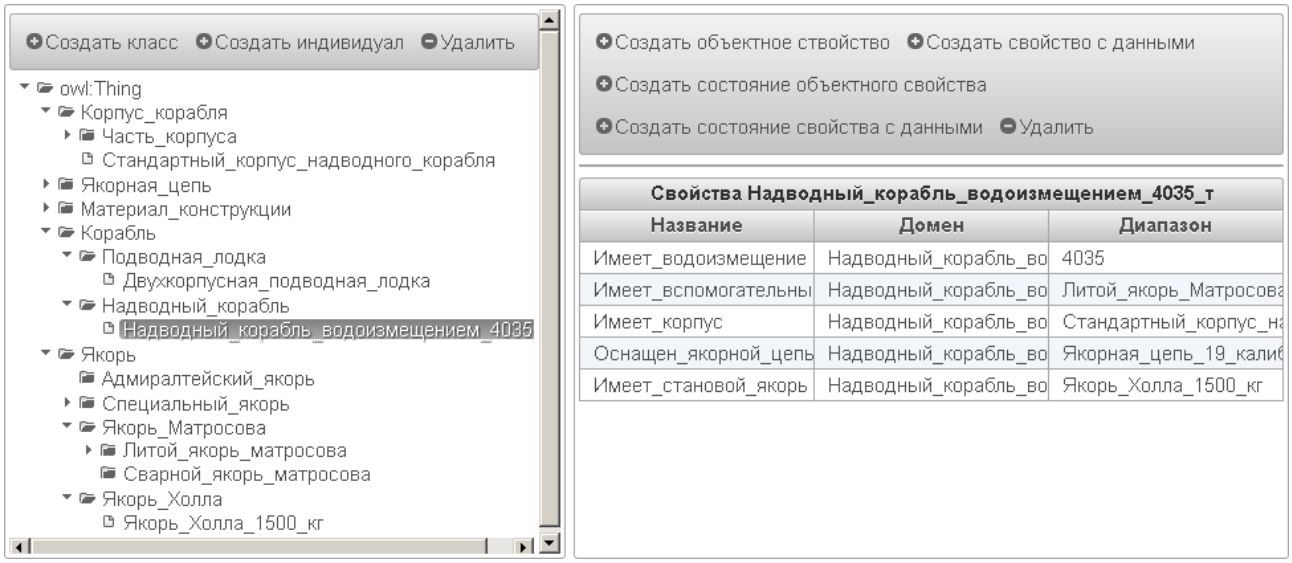

Рисунок 4 – Окно редактора онтологий. Индивидуал «Надводный корабль водоизмещением 4035 тонн» и его состояния

Используемая в данном примере предметная онтология была построена экспертом на основе специализированного справочника боцмана и включает в себя порядка 30 классов, 93 отношения и свыше 120 объектов классов.

Заключение

Таким образом, основное преимущество описанной платформы — наличие единого технологического и информационного пространства процесса принятия решения, информационного поиска с помощью интеллектуального анализа данных, обеспечивающих возможность использования накопленных экспертных знаний о специфике предметной области организации для решения самых различных задач.

Основными направлениями развития рассмотренной платформы являются:

- наращивание функционала путем добавления новых механизмов проведения интеллектуального анализа данных;
- адаптация и унификация ранее реализованных алгоритмов интеллектуального анализа под особенности разработанной платформы с целью их последующего совместного применения при решении широкого круга сложных задач;
- развитие платформы для последующего использования в различных научно-исследовательских проектах;
- интеграция с существующими средами и платформами хранения и обработки данных и знаний, использующихся в производстве и других сферах деятельности.
- адаптация платформы под задачи анализа больших объемов данных (Big Data).

Библиографический список

[Добров и др., 2006] Добров Б.В., Лукашевич Н.В., Лингвистическая онтология по естественным наукам и технологиям: основные принципы разработки и текущее состояние // Десятая национальная конференция по искусственному интеллекту с международным участием (Обнинск, 25-28 сентября 2006 г.) –М.: Физматлит, 2006.

[Гаврилова и др., 2000] Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. –СПб. : Питер, 2000. – 384 с.

[Норенков, 2009] Норенков И.П. Основы автоматизированного проектирования. М: МГТУ имени Баумана, 2009.

[Титов, 2009] Титов Ю. А. САПР технологических процессов. Ульяновск, 2009.

[Wilkins, 2008] Greg Wilkins Jetty vs Tomcat: A Comparative Analysis, 2008. URL – http://www.webtide.com/choose/jetty.jsp.

[Lewis et al, 2014] James Lewis, Martin Fowler Microservicesa definition of this new architectural term, 2014. URL – http://martinfowler.com/articles/microservices.html.

[Наместников и др., 2014] Наместников А.М., Субхангулов Р.А. Онтологически-ориентированная модель классификаций текстовых документов // Материалы IVмеждународной научно-технической конференции OSTIS-2014, Минск. – 2014. С. 385–390.

UNIFORM ONTOLOGICAL DATA MINING PLATFORM

Filippov A.A.,Moshkin V.S., Shalaev D.O., Yarushkina N.G.

Ulyanovsk State Technical University, Russian Federation

al.filippov@ulstu.ru

PostForVadim@yandex.ru

do.shalaev@ulstu.ru

jng@ulstu.ru

Methodology of development, basic principles of construction and architecture of a uniform data mining platform based on ontology describe in this article. Also process of building ontology using the built-in universal platform editor of the example ontology "Handbook boatswain" describes.

Introduction

The desire to automate the process of data specific to the subject area is the need for a single universal data mining tools, to solve various problems, based on the accumulated knowledge and experience of experts in the field:

• creation of the expert system providing an output of recommendations during acceptance of administrative decisions;
• simulation and prediction of indexes of activities of the company according to the previous periods of the reporting;
• search of necessary documentation in volume archive of the organization taking into account its specifics etc.

In this regard, now the task of support of specialists of the organizations of different profiles the universal tools allowing to solve objectives taking into account specifics of specific data domain is actual.

Main Part

The server part of a platform is realized in the Java programming language with use of a framework of Spring which gives ample opportunities for facilitation of development of corporate applications.

The interface of a platform is realized with application of a framework of Google Web Toolkit (GWT) which allows to create RIA (Rich Internet Application) of an Ajax-application on the basis of a Java-code by its compilation in a JavaScript-code.

Use of GWT allows to use Java language by development that reduces quantity of potential errors and does possible debugging of a code of the interface part of a platform. Thus on the client this code is executed in the form of a light-weight JavaScript-code which is supported by all modern browsers.

The principle of microservice architecture is the basis for a platform of intellectual data analysis.

The microservice architecture is a approach in case of which uniform application is distributed and consists of small services which tightly interact among themselves, thus each such service works in own process. For implementation of interaction of services light-weight mechanisms of a remote call of procedures are often used.

Each service realizes specific business process and is executed independent in the isolated container. This approach allows to realize services by means of different programming languages and to use different data stores.

All listed resources, applications and technologies are free and freely spreaded.

Conclusion

The main directions of development of the platform are:

• development of a functionality by adding of new mechanisms of carrying out intellectual data analysis;
• adaptation and unification of earlier realized algorithms of the intellectual analysis under features of the developed platform for the purpose of their subsequent combined use in case of the solution of a wide range of complex challenges;
• development of a platform for the subsequent use in different scientific исследо-вательских projects;
• integration with the existing environments and platforms of storage and data handling and the knowledge which are used in production and other fields of activity.
• adaptation of a platform under tasks of the analysis of large volumes of data (Big Data).