

Национальный исследовательский ядерный университет «МИФИ»

На правах рукописи



Третьяков Евгений Сергеевич

**Методология разработки самоорганизующихся мультиа-
гентных информационно-аналитических систем по
сбору и обработке данных**

Специальность 05.13.01 – Системный анализ, управление и обработка информации
(в информационных системах)

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Москва – 2020

Работа выполнена в Национальном исследовательском ядерном университете «МИФИ»

Научный руководитель:

кандидат технических наук
Алексей Анатольевич Артамонов

Доцент Отделения интеллектуальных кибернетических систем офиса образовательных программ, Национального исследовательского ядерного университета «МИФИ»

Официальные оппоненты:

доктор физико-математических наук, профессор РАН
Арутюн Ишханович Аветисян

Директор Института системного программирования имени Иванникова РАН

доктор технических наук, доцент
Судаков Владимир Анатольевич

Ведущий научный сотрудник Института прикладной математики им. М.В. Келдыша РАН

кандидат физико-математических наук
Зрелов Пётр Валентинович

Международная межправительственная организация Объединенный институт ядерных исследований (ОИЯИ), Начальник научно-технического отдела программного и информационного обеспечения Лаборатории Информационных Технологий (ЛИТ)

Защита диссертации состоится «25» декабря 2020 г. в 15 часов 00 минут на заседании диссертационного совета МИФИ.05.01 федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский ядерный университет «МИФИ» (115409, г. Москва, Каширское шоссе, 31).

С диссертацией можно ознакомиться в библиотеке и на сайте <https://ds.mephi.ru> федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский ядерный университет «МИФИ».

Автореферат разослан «__» _____ 2020 г.

Ученый секретарь

диссертационного совета, д.т.н. Леонова Леонова Наталия Михайловна

ВВЕДЕНИЕ

Исследования и разработки в области информационных систем, использующих агентные технологии, применяются для решения различных задач информационного поиска, при проведении проектных и научных исследований по разработке систем управления и контроля производством и в социальной сфере.

Интеллектуальным агентом или интеллектуальным роботом в данной работе принято считать программно-техническую систему, автономно действующую в некоторой среде (физической или информационной) и решающую задачи, отвечающие интересам (целям) своего владельца (пользователя).

Широкое использование агентных технологий объясняется способностью агентов освобождать пользователей от непосредственного присутствия при решении задач в физической среде и является непосредственным представителем пользователя в информационной среде.

Повышение эффективности агентных технологий осуществляется по двум направлениям: повышение интеллектуального уровня отдельных агентов и повышение интеллектуального уровня управления агентами при коллективном решении поставленных задач.

В качестве основного класса решаемых задач при помощи мульти-агентных технологий в данной работе рассматриваются процессы обработки и анализа данных в информационной среде.

Актуальность исследований. В существующих мультиагентных системах используются агенты, выполняющие разнообразные, но строго регламентированные процессы, а их возбуждение происходит по графику, заданному оператором или по некоторым заранее заданным условиям. Такой подход приводит к высокому уровню простоя вычислительных мощностей при решении задач по подготовке информационно-аналитических материалов в сжатые сроки, т.к. агенты являются недостаточно универсальными, а их побуждение к работе не является событийно-ориентированным.

В диссертации предлагается и реализуется методология группового мультиагентного решения задач конвейерной обработки и анализа интенсивных потоков разнородной информации из различных информационных ресурсов. Групповое автономное решение задач возможно только при наличии автономного управления группой и, в частности, возможности её самоорганизации, т.е. изменения структурных единиц внутри группы и её поведения в зависимости от фактического состояния внешней среды.

Практическое решение проблем самоорганизации в агентных системах представляется наиболее важным направлением повышения их интеллектуальных характеристик и сфер их практического использования.

Поэтому тема данной диссертации представляется **актуальной** как с научной, так и с практической точек зрения.

Цель диссертационной работы состоит в разработке методов и алгоритмов самоорганизации агентов при согласованном решении задач обработки и анализа интенсивных потоков информации.

Для достижения указанной цели **в работе решаются следующие основные задачи.**

1. Изучить особенности агентного поиска информации в глобальной сети и проанализировать подходы к реализации мультиагентных систем с элементами самоорганизации.
2. Исследование и выбор архитектуры мультиагентных информационно-аналитических систем с возможностью самоорганизации, обеспечивающие динамическое взаимодействие с пользователями.
3. Разработка методики предпроектного исследования информационных ресурсов для решения задач целевого сбора информации.
4. Разработка алгоритмов согласованного взаимодействия самоорганизующихся агентов с различными ролевыми предписаниями, позволяющих повысить универсальность системы и гибкость при распределении вычислительных мощностей.
5. Формализация процесса управления реорганизацией агентов, с последующей возможностью проводить мониторинг состояния системы для поддержания ее работоспособности.
6. Разработка и экспериментальное исследование информационно-аналитической системы для обработки потока разнородной неструктурированной информации с элементами самоорганизации.
7. Исследование работоспособности информационно-аналитических систем с элементами самоорганизации в научно-технической и социальной областях.

Объектом исследования являются агентные технологии в тематически ориентированных информационно-аналитических системах.

Предметом исследования являются методы и средства построения мультиагентных систем с элементами самоорганизации.

Методы исследования. В работе используются следующие методы: системного анализа и синтеза сложных систем; автоматической классификации информационных объектов; математической статистики и теории множеств; визуальной компьютерной аналитики.

Научная новизна. Новые научные результаты, полученные лично автором, состоят в следующем:

1. Предложена процедура исследования информационного ресурса для решения задач согласованного сбора информации, состоящая из трех этапов: рекогносцировка, сканирование и инвентаризация. В результате обеспечивается единый подход к разработке агентов.
2. Предложены и обоснованы методы интеллектуализации системы на уровне управления за счет алгоритмов самоорганизации агентов.
3. Разработаны и реализованы алгоритмы контроля выполнения фаз рабочего процесса различными типами агентов с обязательным учетом качественных и временных показателей.
4. Разработан и реализован алгоритм конвейера классификации документов на основе статистических классификаторов первого и второго порядков,

позволяющий проводить тонкую настройку классификации научно-технических текстов в соответствии с требованиями структурных единиц организации-пользователя.

5. Предложены и разработаны методы интеллектуализации агентов для решения информационно-аналитических задач большой размерности на английском и китайском языках.

Обоснованность и достоверность результатов работы подтверждается выводами, полученными в результате экспериментов по использованию агентных технологий при построении мультиагентных информационно-аналитических систем сбора и обработки разнородной информации с элементами самоорганизации и апробацией основных результатов работ на российских и международных конференциях и докладах на научно-технических советах государственных корпораций Ростех и ФГУП «РосРАО» (государственная корпорация Росатом).

Практическая значимость и реализация результатов работы. Теоретические и практические результаты работы использованы при выполнении Государственных заданий Министерства образования и науки Российской Федерации № 2.12611.2018/12.1 «Обеспечение каталогизации и хранения научно-технической информации, полученной из различных неструктурированных источников», № 2.12915.2018/12.1 шифр «Поиск». Программные разработки автора внедрены в Научно-техническом институте межотраслевой информации (НТИМИ) для решения задач по сбору, обработке и визуализации разнородной научно-технической информации для заинтересованных отраслевых организаций и на кафедре «Физические проблемы материаловедения», №9 НИЯУ МИФИ для построения отраслевой информационной системы по ядерным материалам. Автором совместно с экспертами Международного агентства по атомной энергии (МАГАТЭ) отдела Развития ядерной инфраструктуры (Nuclear Infrastructure Development Section) создана и эксплуатируется по всему миру база данных «Nuclear Infrastructure Competency Framework». Автором совместно с экспертами из Мичиганского университета в рамках работы по оптимизации сетевой нагрузки во Всемирной вычислительной сети Большого адронного коллайдера (Worldwide LHC Computing Grid) разработана система анализа и визуализации сетевой топологии. Автором создан учебный курс по подготовке операторов агентных систем для студентов НИЯУ МИФИ и специалистов профильных организаций, использующих мультиагентные системы в своей деятельности.

Апробация результатов работы

Основные результаты исследований работы представлены на следующих конференциях:

1. School on Nuclear Electronics & Computing 2015 (NEC'2015) based on XXV International Symposium on Nuclear Electronics & Computing / Черногория, Будва, Сентябрь 2015.

2. Международная конференция-школа «Проблемы обработки, анализа и управления большими данными в распределенной гетерогенной компьютерной среде для высокоинтенсивных областей науки и бизнес приложений» / Томск, Декабрь 2016.

3. 1st International Early Research Career Enhancement School on Biologically Inspired Cognitive Architectures and Cybersecurity (FIERCES) / Москва, Август 2017.

4. VIII Международная конференция-школа «Распределенные вычисления и GRID-технологии в науке и образовании» / Дубна, Сентябрь 2018.

5. Annual International Conference on Biologically Inspired Cognitive Architectures, the 10th Annual Meeting of BICA Society / США, Сиэтл, Август 2019.

Публикация результатов

Основные положения диссертации опубликованы в 11 печатных работах в изданиях, включенных в список рекомендуемых ВАК, 10 из которых опубликованы в изданиях, индексируемых базой данных Scopus и Web of Science. По научно-техническим разработкам в составе коллектива авторов получено 2 свидетельства о регистрации баз данных и 2 свидетельства о регистрации программы для ЭВМ в Федеральной службе по интеллектуальной собственности.

Основные положения, выносимые на защиту:

1. Процедура исследования информационного ресурса для решения задач согласованного сбора информации.

2. Методы интеллектуализации системы на уровне управления за счет применения алгоритмов самоорганизации агентов.

3. Алгоритмы контроля выполнения рабочего процесса различными типами агентов, обеспечивающие мониторинг полноты обработки информационного потока.

4. Алгоритм конвейера классификации документов на основе статистических классификаторов первого и второго порядков, обеспечивающие распределение информации среди пользователей системы.

5. Методы интеллектуализации агентов для решения информационно-аналитических задач большой размерности на английском и китайском языках.

Личный вклад автора. Основные научные результаты, полученные лично автором, заключаются в разработке методов и алгоритмов интеллектуализации агентов в части построения самоорганизующихся мультиагентных систем. Автором предложена типизация агентов в соответствии с их ролями и их практическая реализация; разработка методов контроля выполнения фаз рабочего процесса; предложены средства интеллектуализации поведения отдельных агентов при обработке Big Data, как на алфавитных, так и на иероглифических языках. Автором лично проведена разработка архитектуры мультиагентной системы, внедренной в Научно-техническом институте межотраслевой информации, в том числе проведены экспериментальные исследования эффективности поведения агентов в самоорганизующихся мультиагентных системах.

Структура и объем диссертационной работы. Диссертационная работа состоит из введения, четырёх разделов, заключения, списка литературы и 2-х приложений. В приложениях приведены свидетельства о Государственной регистрации баз данных и программ для ЭВМ. Общий объем работы составляет 119 страниц (без учёта приложений). Работа содержит 18 иллюстраций, 11 фрагментов кода и 5 таблиц. Список литературы состоит из 105 наименований.

Содержание диссертации и публикаций автора относятся к научной специальности 05.13.01 – системный анализ, управление и обработка информации (в информационных системах) и соответствует следующим позициям паспорта данной специальности:

- формализация и постановка задач системного анализа, оптимизации, управления, принятия решений и обработки информации, п. 2 паспорта;
- разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений и обработки информации, п. 4 паспорта;
- разработка специального математического и алгоритмического обеспечения систем анализа, оптимизации, управления, принятия решений и обработки информации, п. 5 паспорта;
- визуализация, трансформация и анализ информации на основе компьютерных методов обработки информации, п. 12 паспорта.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность темы диссертационной работы, выделяются объект и предмет исследования, определяются цели и задачи работы, приводится список используемых методов исследования, даются сведения о структуре и объеме диссертационной работы.

В первом разделе автор рассматривает международные стандарты в области агентных технологий, вариативность типов информационных источников в глобальной сети, этические вопросы сбора данных и примеры объектов интереса, возникающие в ходе выполнения работ по ситуационному экспресс анализу.

Автор рассматривает существующие определения термина *агент*, которое эволюционировало от определения *человека*, наделённого рядом конкретных функций и полномочий при принятии управленческих решений, до определения *программного* и *интеллектуального агента*, в которые вкладывают понятия компьютерной программы с такими дополнительными свойствами как *автономность*, *социальность*, *реактивность*, *проактивность*.

Рассмотрены международные стандарты построения мульти-агентных систем (МАС), а именно стандарты организации FIPA (*Foundation for Intelligent Physical Agents*). Выделены стандарты, которые являются обязательными для построения FIPA-совместимой МАС: «Стандарт управления агентами» (**SC00023** – «*FIPA Agent Management Specification*») и «Стандарт структуры сообщений на языке общения агентов (ACL)» (**SC00061** – «*FIPA ACL Message*

Structure Specification»). Изучена эталонная модель управления агентами FIPA (см. рис. 1.), включающая в себя такие структурные единицы как *Агент*, *Служба каталогов (DF)*, *Система управления агентами (AMS)*, *Система обмена сообщениями (MTS)*, *Агентная платформа (AP)*, *Программное обеспечение*. В качестве примера приводятся реализации программного обеспечения, которые используют стандарты и протоколы FIPA: полностью FIPA-совместимая платформа для построения интеллектуальных агентов реализованная на языке программирования Java – **JADE** («*Java Agent Development Framework*») (релиз последней версии 2017-06-08) и Агентная платформа **SPADE** («*Smart Python Agent Development Environment*») (релиз последней версии 2018-10-24), использующая отдельные протоколы коммуникации FIPA.

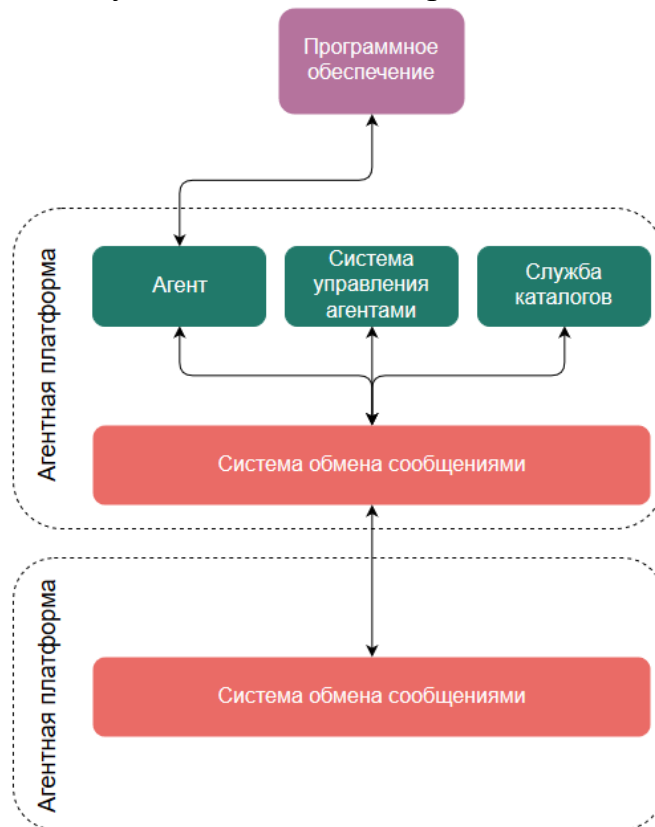


Рис.1. Эталонная модель управления агентами FIPA

При исследовании особенностей агентного сбора информации автором перечислены проблемы при взаимодействии с информационными источниками, в том числе рассмотрены проблемы доступа, сбора и использования информации из глобальной сети с юридической точки зрения. Приведены примеры судебных практик в Соединённых Штатах Америки и Российской Федерации, а также международные практики использования стандартов для осведомления агентов о доступе к разделам информационного источника, а именно – «Стандарт исключений для роботов».

При исследовании публикаций по тематике проектирования самоорганизующихся мультиагентных систем автор особое внимание уделил определению термина «самоорганизация», изучены труды Р. Декарта, У. Эшби, Х. фон Фёр-

стера, П. Гленсдорфа и И. Пригожина. Обзор отечественных и зарубежных источников по агентным самоорганизующимся системам выявил, что проблема создания мультиагентных самоорганизующихся систем рассматривается различными авторами по-своему. Единственно общее положение, на котором прямо или косвенно сходятся все авторы, состоит в том, что самоорганизация является ключевым аспектом мультиагентных систем нового поколения.

Во втором разделе приведена разработанная автором методология построения МАС с элементами самоорганизации, её архитектурная и функциональная модель. Автором предложена методика системного анализа информационной среды, состоящей из трех этапов: *Рекогносцировка*, *Сканирование* и *Инвентаризация*. Методика регламентирует стадии изучения информационного ресурса для построения агентов.

В общем виде самоорганизующуюся мультиагентную систему можно представить в следующем виде

$$MAS\{E, O, A, W, R, I\}, \text{ где} \quad (1)$$

E – среды, в которых работает МАС,

O – объекты интереса пользователя,

A – интеллектуальные агенты,

W – рабочие процессы,

R – отношения между сущностями в мультиагентной системе (агент-агент, агент-среда и т.д.),

I – агентные предписания.

Максимальное количество агентов в мультиагентных системах устанавливается разработчиком, на основе возможностей аппаратной платформы.

«Рабочий процесс», как правило, состоит из фаз, выполнение которых контролируется «Агентами-лидерами». Каждая фаза начинается с проверки необходимых данных или статуса уже выполненных заданий на предыдущей фазе и заканчивается генерацией новых заданий или объявлением окончания «Рабочего процесса». На каждой фазе «Агент-лидер» набирает разных «Агентов-исполнителей» в зависимости от поставленной конечной цели и формирует «Агентную рабочую группу».

Рабочий процесс является сложной сущностью, определяющей необходимые стадии в мультиагентной системе для достижения поставленных целей.

Рабочий процесс формализован в следующем виде:

$$W = \{St_1 R_1, \dots, St_N R_N\}, \quad (2)$$

$$N \in \mathbb{N} \wedge N \neq 0, \text{ где}$$

St – стадии рабочего процесса,

R – отношения между стадиями и агентами,

N – количество стадий рабочего процесса,

$St_i R_i$ – это пара элементов, состоящая из стадии рабочего процесса (St_i) и отношения между ними (R_i), где $i = \overline{1, N}$.

Для успешного выполнения стадии рабочего процесса может использоваться как один агент, так и агентная группа. В каждой фазе могут быть выделены критические и некритические задания.

Автором предложена архитектура МАС (см. рис. 2), включающая такие компоненты как: *Веб-платформа*, *Хранилище*, *Система управления агентами* и *Система коммуникации*.

В предложенной архитектуре агенты находятся в веб-платформе, что обуславливается необходимостью взаимодействия агентов как с хранилищем, посредством включённого программного обеспечения веб-платформы, так и взаимодействием агентов с пользователем, где агент может быть активизирован по запросу пользователя.

Агенты разделены на внешний и внутренний контур по принципу среды их работы и сферы взаимодействия: *внутренние агенты* не обладают функцией выхода в глобальную сеть, но располагают большими полномочиями во внутренней сети, а *внешние* – наоборот. Введены понятия агентных предписаний: *поисковых* и *операционных*.

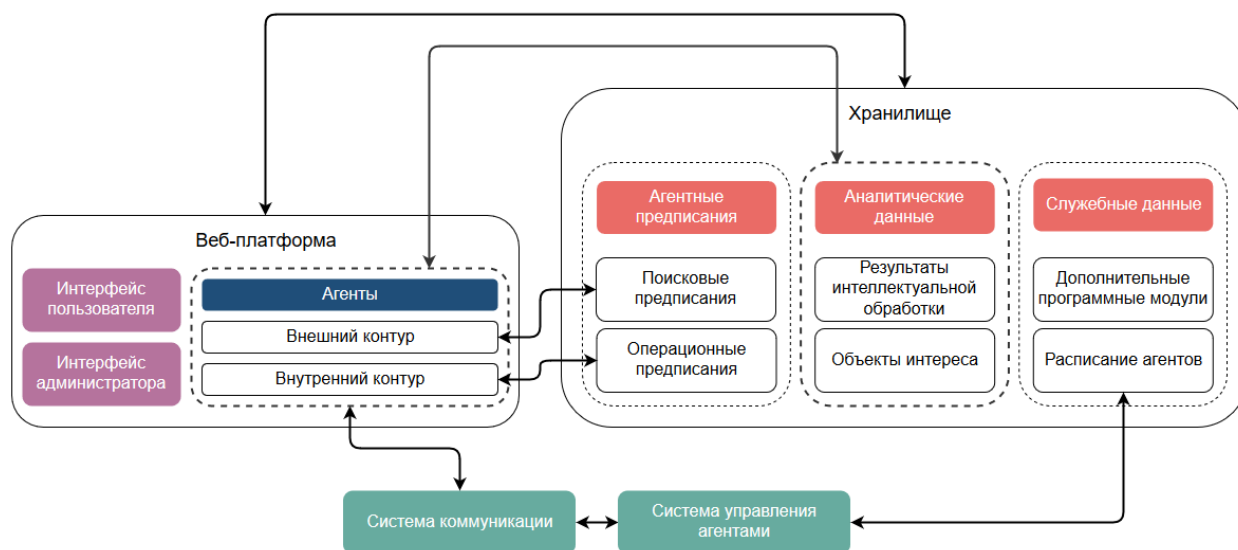


Рис.2. Архитектура МАС

Автором разработана система самоорганизации агентов (см. рис. 3), которая позволяет решать задания в рамках рабочего процесса в автоматизированной форме.

«Агент-планировщик», согласно установленному расписанию, сканирует объекты «Информационный источник» и создаёт в базе данных объекты «Рабочий процесс». На основе предписаний из записей «Информационный источник», «Агенты-лидеры» создают задания для «Агентов-исполнителей», тем самым организуя «Агентные рабочие группы».

На примере решения задачи сбора и обработки информации из глобальной сети дана классификация агентов в рамках автономного группового управления: «Агент-планировщик», «Агент-лидер» и «Агент-исполнитель», введены понятия сложных объектов «Информационный источник», «Рабочий процесс» и «Агентная рабочая группа». Записи «Рабочий процесс» используются «Агентами-лидерами» для координации работы «Агентной рабочей

группы». Получение сведений от «Агентов-исполнителей» осуществляется через диалог с «Агентами-лидерами», после чего на основе данных из «Рабочего процесса» ведётся координация дальнейших работ.

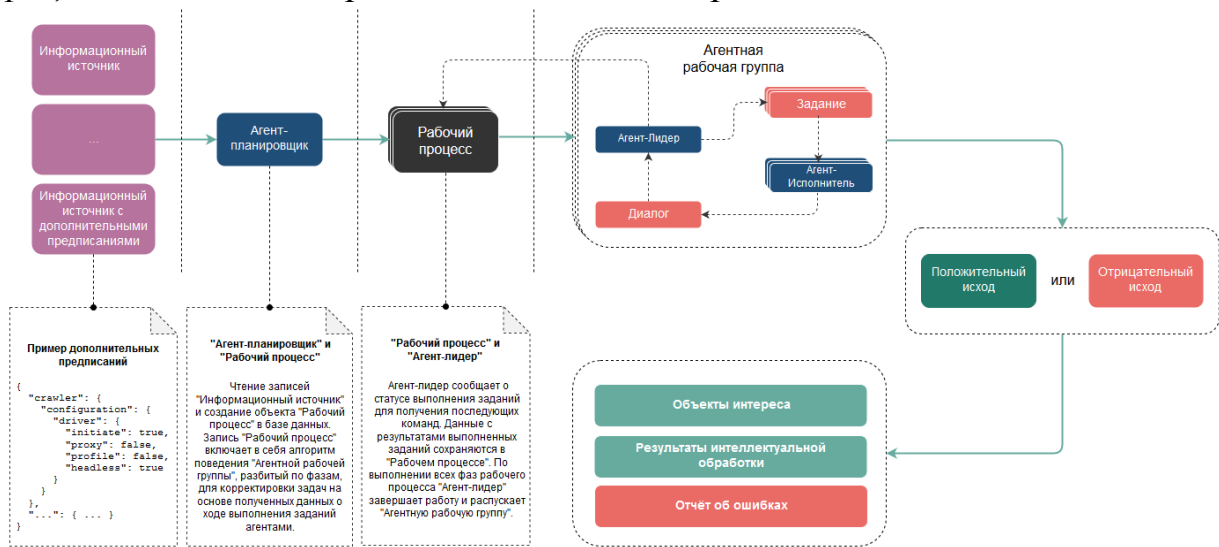


Рис.3. Схема самоорганизации агентов в МАС

В рамках агентной рабочей группы реализуются алгоритмы самоорганизации агентов.

Самоорганизация агентов – это автономное формирование алгоритмов группового управления агентами.

Алгоритмы самоорганизации агентов опираются на рабочий процесс, процессы обмена знаниями и процессы принятия решений агентами. На Рис. 4 представлена типовая схема процесса самоорганизации агентов при выполнении рабочего процесса. Цифрами на рисунке обозначена последовательность действия «Агента-лидера».

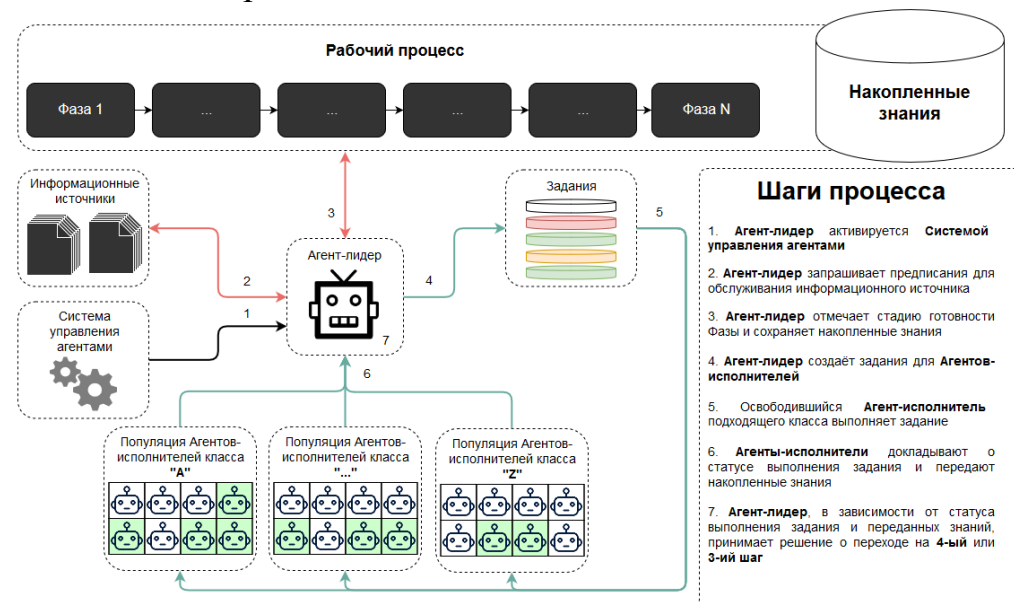


Рис.4. Схема процесса самоорганизации задания агентов при выполнении рабочего процесса

Обмен сообщениями между агентами может вестись в свободной форме с указанием методов для обработки данных на том или ином этапе или при помощи адаптации международных стандартов, таких как FIPA.

В предложенной автором функциональной модели самоорганизации автономные решения, как элементы самоорганизации, принимаются в следующих случаях: формирование «Агентом-лидером» «Агентной рабочей группы»; формирование «Агентом-лидером» группы «Агентов-исполнителей» для выполнения всех поисковых предписаний; распределение «Агентом-лидером» заданий на работу «Агентов-исполнителей» с определенными источниками информации; принятие «Агентом-лидером» решения о критической или некритической ошибке в работе «Агентов-исполнителей» и, соответственно, принятия решения об остановке или продолжении агентных поисковых процедур.

На основе результатов исследования информационной среды, разработаны методы согласованного взаимодействия агентов внешнего и внутреннего контуров, приведены примеры разработки алгоритмов поведения для агентов и конкретные примеры формы представления поисковых и операционных предписаний.

В третьем разделе рассмотрено применение «Методологии построения мультиагентных систем» для инженерной реализации решения задач обработки интенсивных потоков научно-технической информации по широкому спектру тематических направлений. Данный пример характерен тем, что задействует множество агентов разного типа для асинхронной обработки информационных материалов, а также раскрывает возможность построения распределённой системы вычислений с включением стороннего ПО, как элемента МАС.

Назначение МАС – обработка неструктурированного потока информации, представленного в виде постоянно пополняющихся информационных материалов в сетевом хранилище, для дальнейшего анализа экспертом в определённой научной области.

Информационный материал (ИМ) – это директория в сетевом хранилище, располагающая неопределённым количеством файлов разного формата, например, файлы с расширением pdf, rtf, odf и форматов MS Office, таких как doc(x), xls(x), ppt(x), vsd(x).

Обработка информационного материала – это процесс прохождения ИМ через все стадии агентной интеллектуальной обработки, описанные в предписаниях, как алгоритм поведения «Агента-лидера».

Алгоритм поведения «Агента-лидера» разбит на семь фаз и включает следующие стадии обработки ИМ:

1. *Поиск* – поиск ИМ в сетевом хранилище.
2. *Регистрация* – регистрация файлов ИМ на обработку.
3. *Извлечение* – извлечение текста из файлов ИМ.
4. *Классификация* – классификация файлов по научному направлению.
5. *Библиография* – извлечение библиографических данных.

6. *Квазиреферат и фактография* – составление квазиреферата и извлечение фактографии.

7. *Заключение* – написание заключения по ИМ на основе содержащихся файлов.

Разделение стадий обработки ИМ на фазы позволяет «Агенту-лидеру» контролировать асинхронное выполнение задач в рамках рабочего процесса.

Гибкость МАС обеспечивается настройкой алгоритма поведения «Агента-лидера» файлом в формате JSON, который относится к каждой занесённой записи об информационном источнике. Таким образом, алгоритм поведения «Агента-лидера» зависит от источника данных. На рис. 5 изображён алгоритм контроля выполнения рабочего процесса «Агентом-лидером» с первой по четвёртую фазу. Создание задач при прохождении каждой фазы сопровождается предоставлением необходимых данных будущему «Агенту-исполнителю».

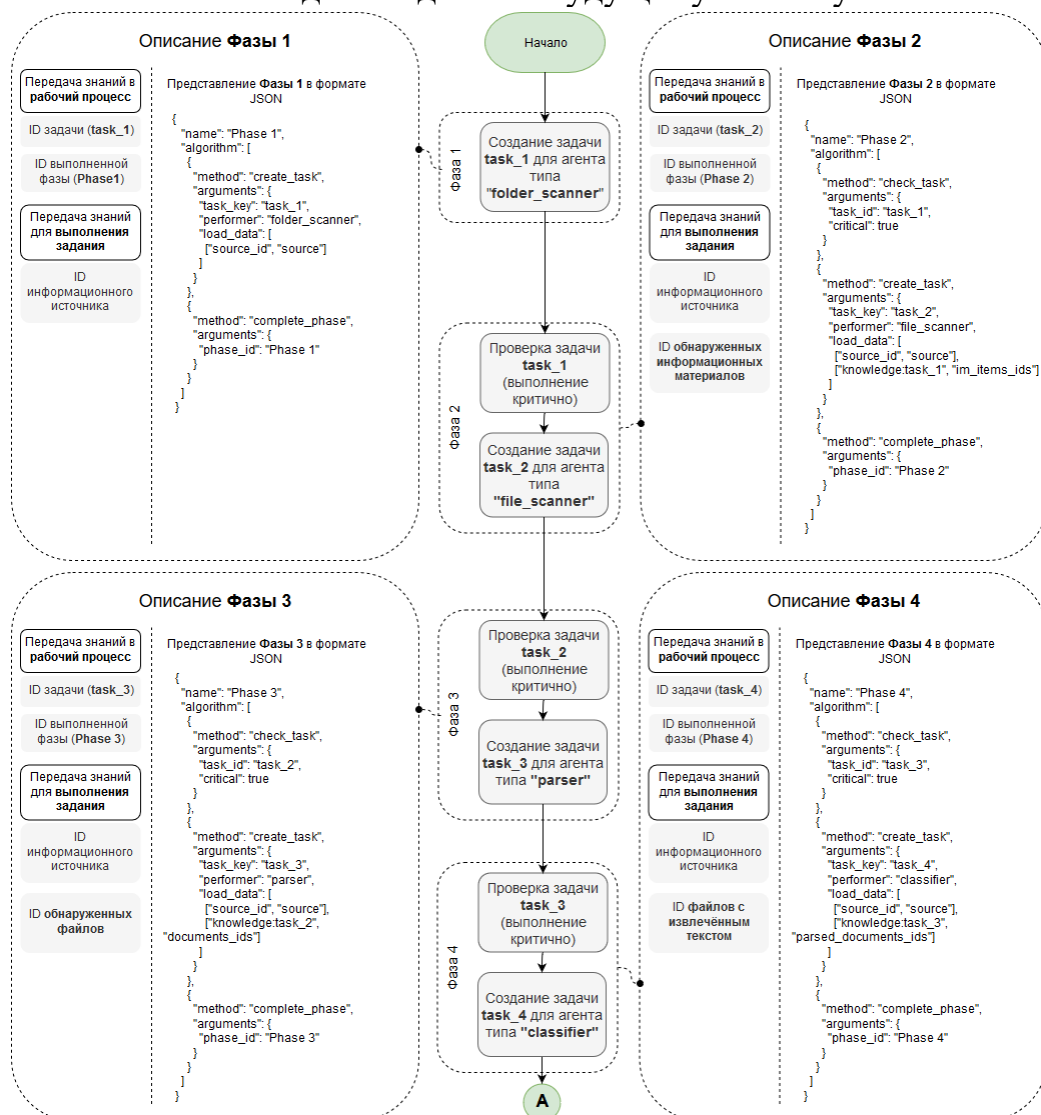


Рис. 5. Алгоритм контроля выполнения рабочего процесса «Агентом-лидером» с первой по четвёртую фазу

Общее поведение «Агентов-исполнителей» описывает их жизненный цикл и базовые функции. Жизненный цикл «Агентов-исполнителей» состоит из

трёх стадий: инициализация, выполнение поставленной задачи и вывод «Агента-исполнителя» из «Агентной рабочей группы».

Решение об инициализации «Агента-исполнителя» зависит от ограничений по популяции каждого типа «Агента-исполнителя», описанных в настройках системы.

Выполнение поставленной задачи осуществляется посредством базовых методов, таких как «*process*» и «*save_knowledge*». Метод «*process*» позволяет воспроизводить алгоритм обработки данных, который описан в предписаниях источника информации, а метод «*save_knowledge*» позволяет сохранять полученные знания для последующего выполнения работ другими агентами в рамках «Агентной рабочей группы».

Вывод «Агента-исполнителя» из «Агентной рабочей группы» характеризуется освобождением занятых вычислительных ресурсов и уведомлением о своём выходе всех «Агентов-исполнителей» в рамках «Агентной рабочей группы».

Частные поведения «Агентов-исполнителей» – это те функции, которые отличают «Агентов-исполнителей» одного типа от «Агентов-исполнителей» других типов. Автором описаны частные поведения «Агентов-исполнителей» следующих типов: «*folder_scanner*», «*file_scanner*», «*parser*», «*classifier*», «*publication_extractor*», «*document_analyser*» и «*im_analyser*». Данные типы агентов отвечают за установленные семь стадий обработки ИМ соответственно.

Подробно рассмотрен интеллектуальный агент типа «*classifier*», поведение которого заключается в классификации файлов ИМ. Классификация файлов ИМ производится по введённым в систему Конвейерам классификаций, которые могут включать статистические классификаторы первого (СКП) и второго (СКВ) уровня.

Конвейеры классификации описывают порядок применения СКП и СКВ для определения конечной рубрики. Таким образом возможно определять рубрики не только по широким научным направлениям, но и по узкоспециализированным направлениям в рамках научного института, коллектива или отдельного эксперта путём обучения на предоставляемых массивах документов или при непосредственном участии экспертов.

СКП включают в себя термины и рубрики, где каждый термин имеет отношение к рубрике с определённым весом. Отношения всех рассматриваемых терминов к рубрикам задаётся в виде такой матрицы инцидентности A , элементы которой a_{ij} определяют веса терминов к рубрикам. Для более подробного рассмотрения введено определение гиперграфа:

$$\begin{aligned}
R &= \{r_1, \dots, r_N\} - \text{множество рубрик СКП,} \\
T &= \{t_1, \dots, t_M\} - \text{множество терминов СКП,} \\
r_i &= \{t_{i_1}, \dots, t_{i_k}\}, \text{ где } i \in \mathbb{N}, t_{i_j} \in T, j = \overline{1, k} \text{ и}
\end{aligned} \tag{3}$$

$$\forall t_{i_j}, t_{i_r}, \text{ где } j \neq r, t_{i_j} \neq t_{i_r}$$

В таком случае матрица инцидентности $\|A\|$ описывается следующим образом:

$$A: (N \times M), a_{i,j} \in [0,1] \tag{4}$$

При этом поступающий текст из слов $S = \{s_1, \dots, s_L\}$, где $L \in \mathbb{N}$ и s_i – слово на естественном языке, может быть классифицирован если $S \cap T \notin \emptyset$, т.е. $\exists s_i = t_j \in T$.

Определение рубрики текста осуществляется посредством вычисления упорядоченного набора значений $\{b_i\}$, где $i = \overline{1, M}$, определяющих количество вхождений термина t_i в S (если термин t_i не встречался в тексте S , то $b_i = 0$).

Тогда программную функцию F , получающую на вход параметры S, T, A, R и передающую на свой выход рубрику r , где $r \in R$, т.е. определение рубрики классификатором можно описать следующим образом:

$$F(S, T, A, R) \rightarrow r \in R$$

$$r = r_{i_{\max}} \text{ такая что } c_{i_{\max}} = \max\{c_i\}, \text{ где} \tag{5}$$

$$c_i = \sum_{j=1}^M a_{ij} (1 - e^{-b_j}) \sqrt{\pi}, i \in \overline{1, N}$$

Статистический классификатор второго (СКВ) включает в себя отношения рубрик одного рубрикатора и рубрик другого рубрикатора с указанием весовых характеристик.

Отношения одних рубрик к другим задаётся в виде такой матрицы инцидентности $\hat{A}: (U \times W)$, что $\hat{a}_{ij} \in [0,1]$, где $\hat{a}_{ij}$ – коэффициент, задающий вес отношения рубрики r_i к рубрике $\hat{r}_j$, где множества рубрик R и $\hat{R}$ описываются следующим образом:

$$\begin{aligned}
R &= \{r_1, \dots, r_U\}, \hat{R} = \{\hat{r}_1, \dots, \hat{r}_W\} \\
r_i &= \{\hat{r}_{i_1}, \dots, \hat{r}_{i_k}\}, \text{ где } i \in \mathbb{N}, \hat{r}_{i_j} \in \hat{R}, j = \overline{1, k} \text{ и}
\end{aligned} \tag{6}$$

$$\forall \hat{r}_{i_j}, \hat{r}_{i_r}, \text{ где } j \neq r$$

$$\hat{r}_{i_j} \neq \hat{r}_{i_r}$$

Вычисление рубрики по СКВ осуществляется посредством использования множества $\{c_i\}$ при вычислении рубрики по СКП, при этом рубрика СКВ может быть определена если $\sum_{i=1}^U c_i \neq 0$, $i \in \overline{1, U}$.

Таким образом функция классификатора, определяющая рубрику по СКВ, описывается следующим образом:

$$\begin{aligned} \hat{F}(c_i, R, \hat{A}, \hat{R}) &\rightarrow \hat{r} \in \hat{R} \\ \hat{r} = \hat{r}_{i_{max}} | d_{i_{max}} &= \max\{d_i\}, \text{ где} \\ d_i &= \sum_{j=1}^W c_i \hat{a}_{ij}, \forall i \in \overline{1, U} \end{aligned} \quad (7)$$

В различных *Конвейерах классификации* могут быть использованы на разных стадиях одни и те же СКП и СКВ, поэтому в арсенале «Агента-исполнителя» типа «*classifier*» предусмотрен механизм кэширования, т.е. механизм запоминания промежуточных вычислений с целью оптимизации затрат вычислительных мощностей.

Приведены данные об экспериментальном исследовании поведении агентов в МАС при решении задачи обработки поступающих ИМ, как на контрольной выборке, так и в круглосуточном автоматическом режиме. Контрольная выборка состояла из 255 ИМ (1 285 текстовых файлов различного формата общим объёмом в 1.5 Гб).

Платформой для развёртывания МАС послужила виртуальная машина с операционной системой на базе ядра Linux 4.5, 8 192Mb RAM и 4-мя ядрами CPU.

Среднее количество файлов в одном ИМ составило 5 единиц. Количество затраченного времени на обработку всего массива составило 4 963 секунды (83 минуты), при этом среднее количество затраченного времени на один ИМ составило 19 секунд, а на обработку одного документа – 4 секунды. Среднее количество времени, затраченное «Агентами-исполнителями» на обработку заданий представлено в таб. 1.

Таблица 1

Показатели среднего затраченного времени по типам «Агентов-исполнителей»

Тип «Агента-исполнителя»	Среднее затраченное время на выполнение одного задания (в секундах)
folder_scanner	0.1
file_scanner	0.2
parser	90.5
classifier	13.7
publication_extractor	9.3
document_analyser	35.0
im_analyser	1.1

Показатели среднего затраченного времени при использовании агентных технологий в поставленном эксперименте демонстрируют практическую применимость инструментов МАС для решения аналитических задач большой размерности (Big Data).

Исследованы возможности реорганизации агентов в рамках выполнения рабочего процесса при обработке научно-технического потока информации. Способность к реорганизации агентных рабочих групп демонстрирует интеллектуальность системы на уровне управления, позволяющая выгодно распределять вычислительные мощности при различных рабочих процессах.

В четвёртом разделе рассмотрена возможность применения агентных технологий для решения прикладных информационно-аналитических задач большой размерности, за счет повышения интеллектуального уровня отдельных агентов. Приведены примеры по формированию единой базы данных атомных электростанций и изучение научно-технического развития Китайской Народной Республики (КНР).

В качестве примера автоматического сбора и актуализации информации поставлена и решена задача поиска информации об атомных электростанциях по всему миру. Выбраны следующие информационные ресурсы: МАГАТЭ (International Atomic Energy Agency) Международное агентство по атомной энергии; совместный проект World Nuclear Industry Status Report & VISIONSCARTO; и WNA (World Nuclear Association) Всемирная ядерная организация, занимающаяся продвижением атомной энергетики. Поля с данными, содержащие описание той или иной атомной электростанции, в рассматриваемых источниках различаются, то есть не имеют строгой структуры и организации данных. Иными словами, в совокупности эти информационные ресурсы дополняют друг друга, предоставляя пользователю наиболее исчерпывающие данные относительно атомных электростанций.

Агентный сбор данных из нескольких информационных источников осуществляется посредством специальных программных инструментов, имеющих структуру, приведенную на рис. 6.

При осуществлении сбора данных из трех выбранных информационных источников получено 789 объектов атомных электростанций (МАГАТЭ - 441; World Nuclear Industry Status Report & VISIONSCARTO - 754; и WNA - 679). С помощью агентных технологий реализовано форматирование и агрегация данных из трёх авторитетных информационных источников и в общей базе данных. Разработанные программные агенты позволяют получать актуальную информацию по заданной тематике, так как обновление данных происходит автоматически по заданному расписанию. Достоверность информации обеспечивается сравнением значений отдельных полей из разных источников.

Для решения задач автоматизированной оценки технологических объектов, автором введено два типа аналитических поисковых индексов: параметрический и семантический. Параметрический поисковый индекс позволяет пользователям найти информацию о технологии или материале с определенными фи-

зическими параметрами, значения которых лежат в заданном интервале, в отличие от традиционного поиска в подстроке, который позволяет пользователям найти только определенное значение.

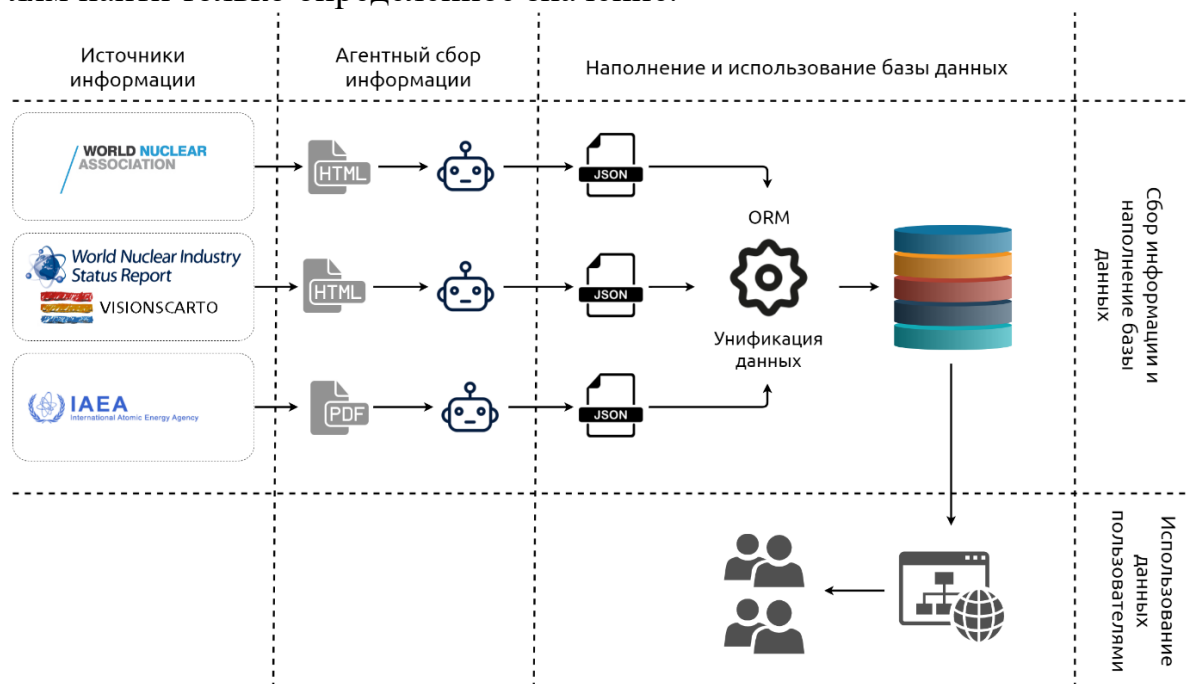


Рис. 6. Архитектура программного обеспечения для агентного сбора данных из нескольких информационных источников

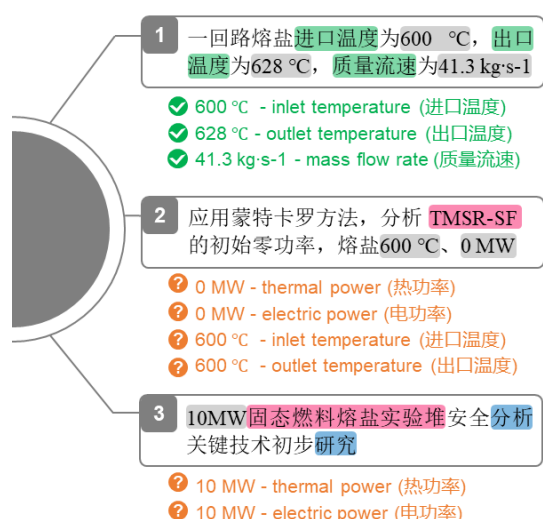
Идея семантического аналитического индекса основана на понятии жизненного цикла технологии, что позволяет определить текущее состояние развития конкретной технологии, на основе разработанной автором модели жизненного цикла.

В работе реализованы алгоритмы индексации для китайского языка, имеющего свои особенности. Были созданы онтологии физических параметров и жизненного цикла. Каждый термин на китайском языке, входящий в онтологии, соотнесен с термином на русском и английском языках.

Семантическая индексация указывает положение «технологического объекта» на разных этапах жизненного цикла. В настоящее время не существует единой унифицированной модели жизненного цикла. Разработана модель жизненного цикла, состоящая из пяти стадий - исследование. Разработка, производство, эксплуатация, утилизация.

Каждая стадия характеризуется набором наиболее часто используемых глаголов, которые вошли в онтологию. Например, глаголы «производить», «строить», «изготавливать» относятся к стадии производства.

На рис. 7 приведены результаты экспериментов по обработке 30 китайских научных статей из открытых источников информации, посвященных ядерной энергетике. Первый блок точно определяет выявленные параметры, во втором блоке содержатся предположения в третьем блоке приведено предложение из семантического индекса.



1 一回路熔盐进口温度为600 °C，出口温度为628 °C，质量流速为41.3 kg·s⁻¹

✓ 600 °C - inlet temperature (进口温度)

✓ 628 °C - outlet temperature (出口温度)

✓ 41.3 kg·s⁻¹ - mass flow rate (质量流速)

2 应用蒙特卡罗方法，分析 TMSR-SF 的初始零功率，熔盐600 °C、0 MW

❓ 0 MW - thermal power (热功率)

❓ 0 MW - electric power (电功率)

❓ 600 °C - inlet temperature (进口温度)

❓ 600 °C - outlet temperature (出口温度)

3 10MW固态燃料熔盐实验堆安全分析关键技术初步研究

❓ 10 MW - thermal power (热功率)

❓ 10 MW - electric power (电功率)

Value	Unit	Parameter (cn)	Parameter (en)	Parameter (ru)	Degree of confidence	Article ID
600	°C	进口温度	inlet temperature	температура на входе	assertion	12
628	°C	出口温度	outlet temperature	температура на выходе	assertion	12
41.3	kg·s ⁻¹	质量流速	mass flow rate	массовый расход	assertion	12
600	°C	进口温度	inlet temperature	температура на входе	assumption	12
600	°C	出口温度	outlet temperature	температура на выходе	assumption	12
0	MW	热功率	thermal power	тепловая мощность	assumption	12
0	MW	电功率	electric power	электрическая мощность	assumption	12

parameter name technology

value and UoM verb

Рис. 7. Фрагмент из таблицы результатов эксперимента

В правой части рисунка находится фрагмент таблицы базы данных, содержащей результаты обработки текста. К записям базы данных можно применять различные фильтры. Например, пользователь базы данных может выбрать определенный параметр и отсортировать записи по значениям этого параметра. Таким образом, можно найти технологию со значениями параметров, лежащими в определенном интервале.

Семантические и параметрические аналитические поисковые индексы могут быть использованы как по отдельности, так и совместно для формирования сложных запросов. Их применение позволяет не только быстро обрабатывать тексты, но и находить технологии с заданными параметрами и на определенном этапе жизненного цикла. Это ускоряет и облегчает процесс анализа и позволяет исследователям более эффективно обрабатывать большие данные

Изучение научно-технического развития КНР осуществлялось посредством анализа опубликованных в апреле 2019 года патентных данных «Департамента разработки техники и вооружения Центрального военного комитета Китая» («The Equipment Development Department of China's Central Military Commission (CMC)»), а именно 7 560 патентов. Доступ к данным такого рода предоставляется специализированными веб-системами, которые обеспечивают пользователя инструментами поиска и фильтрации для нахождения определённого объекта – патента.

Однако, на практике, высокую ценность имеют данные, описывающие современные тенденции, уровень развития технологий по научным направлениям, общее бюджетирование в рамках определённой структуры и т.д.

Представление о данных хорошо формируется посредством построения и изучения графов. Например, при использовании таких атрибутов патента как заявитель (организация) и код IPC, строится граф, где узлами первого типа являются патентные семьи IPC, узлами второго типа — организации, указанные в патентах, а рёбрами — отношения между узлами первого и второго типов.

При этом, рёбра могут быть взвешенными, т.е., в данном случае, характеризовать количество патентов, зарегистрированных организацией по различным рубрикам IPC.

На рис. 8 узлами фиолетового цвета представлены организации-заявители, а бирюзовым — патентные семейства IPC. Для визуализации графа наиболее распространёнными являются алгоритмы, основанные на аналогиях физических принципов притяжения и отталкивания тел или частиц по закону Гука, Кулона и др. один из таких алгоритмов является ForceAtlas 2, который был использован при построении данного графа.

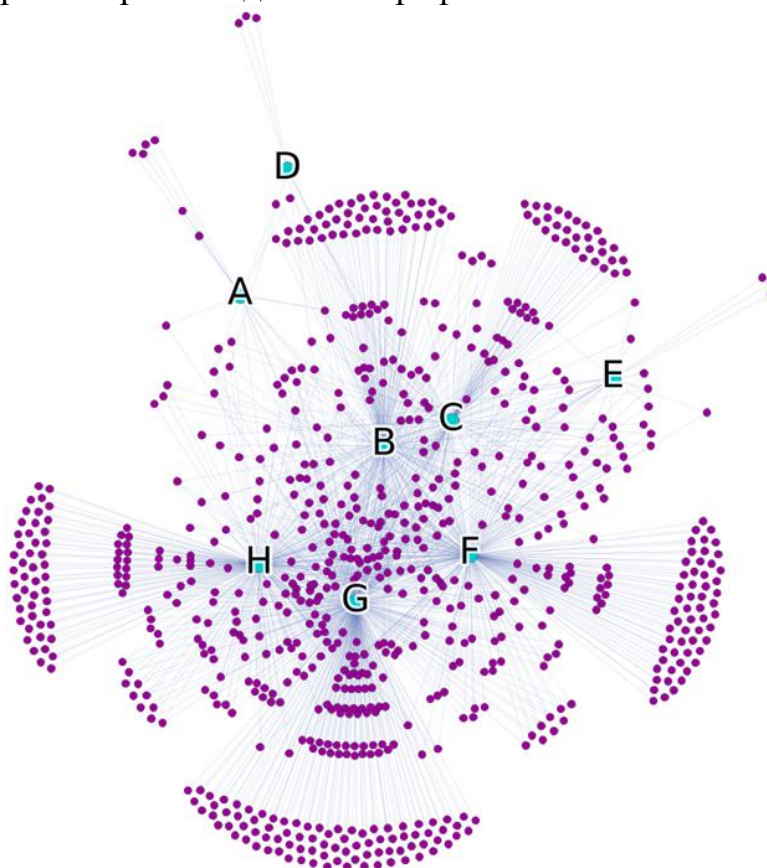


Рис. 8. Граф организации-заявители и патентные семьи IPC

Данный граф чётко разграничивает узко- и широко-специализированные организации, располагая «островками» на внешних «орбитах» графа узкоспециализированные и стягивая к центру графа широко-специализированные организации. Для более детального изучения данных следует рассматривать отдельные фрагменты графа в большем масштабе при помощи специализированных инструментов навигации.

В конце четвёртого раздела приведен разработанный автором учебный курс, рассчитанный на студентов Института международных отношений НИЯУ МИФИ и специалистов профильных организаций использующие МАС в своей деятельности. Каждая тема курса сопровождается практической задачей, что позволяет придерживаться идеологии «Учиться путём практики» или «Обучение в процессе работы».

В заключении отражены основные результаты, полученные в данной работе.

Приложения содержат акты о внедрении результатов диссертационной работы; полученные свидетельства о регистрации баз данных и программ для ЭВМ.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

В диссертации создана методология разработки и проектирования нового класса мультиагентных информационно-аналитических систем по сбору и обработке данных, в которых вводятся интеллектуальные подходы взаимодействия с пользователями. Разработанные на основе данной методологии информационно-аналитические системы позволяют расширить круг пользователей и повысить уровень информационного обслуживания с учётом индивидуальных предпочтений пользователей.

Основные результаты диссертационной работы заключаются в следующем:

1. Разработана методика предпроектного исследования информационных ресурсов для решения задач целевого сбора информации, включающая этапы рекогносцировки, сканирования и инвентаризации. Результатом является формализованная процедура для моделирования и последующей настройки соответствующего агента на информационный ресурс.

2. Предложена и обоснована архитектура мультиагентной информационно-аналитической системы, которая позволяет повысить универсальность системы за счет вводимых различных рабочих процессов и гибкость при распределении вычислительных мощностей за счет своей модульности, которая обеспечивает вертикальное и горизонтальное масштабирование.

3. Применительно к сформированной архитектуре разработаны методы и алгоритмы самоорганизации агентов с различными ролевыми предписаниями. Разработанные алгоритмы используются для обеспечения процесса управления реорганизацией агентов при выполнении ими коллективной работы в решении информационно-аналитических задач.

4. Разработаны методы согласованного взаимодействия агентов внешнего и внутреннего контуров. Агенты внешнего контура ориентированы на взаимодействие с внешней средой для сбора информации. Агенты внутреннего контура обеспечивают обработку и доставку информации конечному пользователю. Функция накопления знаний Агентной рабочей группой позволяют проводить мониторинг состояния системы и поддерживать ее работоспособность.

5. Исследованы и разработаны способы автоматизированной рубрикации потока научно-технической информации. Использование методов математической статистики позволило повысить точность и быстроедействие информационно-аналитической системы.

6. Обоснована и разработана процедура формирования рабочего процесса для управления реорганизацией агентов. Алгоритм контроля выполнения рабочего процесса позволяет операторам системы проводить мониторинг состояния системы и обеспечивать ее рабочее состояние.

7. Предложена и разработана программная платформа для создания информационно-аналитических агентных систем с функциями составления рабочих процессов и их миграции на другие дистрибутивы данной платформы. Функции составления и миграции рабочих процессов позволяют проводить оперативное развертывание системы на новых вычислительных мощностях.

8. Разработан и применен алгоритм новой конвейерной классификации научно-технического потока данных большой размерности. Конвейерная классификация приводится как пример способности информационно-аналитической системы повысить свою интеллектуальность на уровне отдельного агента.

9. Исследованы возможности реорганизации агентов в рамках выполнения рабочего процесса при обработке научно-технического потока информации. Способность к реорганизации агентных рабочих групп демонстрирует интеллектуальность системы на уровне управления, позволяющая выгодно распределять вычислительные мощности при различных рабочих процессах.

10. Разработана и испытана информационно-аналитическая система для обработки потока разнородной неструктурированной научно-технической информации. Экспериментальное исследование информационно-аналитической системы демонстрирует правильность основных теоретических положений, изложенных в диссертации, и универсальность применения методологии разработки самоорганизующихся мультиагентных информационно-аналитических систем.

11. Продемонстрирована практическая значимость использования агентной системы при обработке информации на иероглифических языках на примере сбора и обработки более семи тысяч открытых патентов КНР по технологиям двойного назначения.

12. Составлен и внедрен в учебную практику курс подготовки операторов агентных систем для студентов магистратуры НИЯУ МИФИ и специалистов профильных организаций, использующих мультиагентные системы в своей профессиональной деятельности.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

I. Публикации, представленные в международных базах цитирования Scopus и Web of Science

1. Tretyakov E.S. Algorithmization of search operations in multiagent information-analytical systems / Ananieva A.G., Artamonov A.A., Galin I.U, Tretyakov E.S., Kshnyakov D.O. // Journal of Theoretical and Applied Information Technology, Nov. 2015, Vol. 81, No. 1. - pp. 11-17.

2. Tretyakov E.S. The express case analysis results visualization / Artamonov A.A., Ananieva A.G., Tretyakov E.S., Onykiy B.N., Pronicheva L.V., Ionkina K.V., Suslina A.S. // *Scientific Visualization*, 2016, Vol. 8, No.3. - pp 25-34.

3. Tretyakov E.S. A three-tier model for structuring of scientific and technical information / Artamonov A.A., Ananieva A.G., Tretyakov E.S., Kshnyakov D.O., Onykiy B.N., Pronicheva L.V. // *Journal of Digital Information Management*, June 2016. Vol.14. No.3. - pp 184-193.

4. Tretyakov E.S. Agent Technologies for Polythematic Organizations Information-Analytical Support / Onykiy B.N., Artamonov A.A., Ananieva A.G., Tretyakov E.S., Pronicheva L.V., Ionkina K.V., Suslina A.S. // *Procedia Computer Science*, 2016, Vol. 88, pp. 336-340.

5. Tretyakov E.S. Methodology of Learning Curve Analysis for Development of Incoming Material Clustering Neural Network // Onykiy B.N., Tretyakov E.S., Pronicheva L.V., Galin I.U., Ionkina K.V., Cherkasskiy A.I. // *BICA 2017: Biologically Inspired Cognitive Architectures (BICA) for Young Scientists*, Jul. 2017, pp. 133-138.

6. Tretyakov E.S. Visualization of large samples of unstructured information on the basis of specialized thesauruses / Artamonov A.A., Tretyakov E.S., Ionkina K.V., // *Scientific Visualization*, Dec. 2017, Vol. 9, No. 5, pp. 54-58.

7. Tretyakov E.S. Electronic document processing operating map development for the implementation of the data management system in a scientific organization / Artamonov A.A., Ionkina K.V., Tretyakov E.S., Timofeev A.I. // *Procedia Computer Science*, 2018, Vol. 145, pp. 248-253.

8. Tretyakov E.S. Agent technology situational express analysis in assessment of technological development level of the BRICS countries / Koshlan D.I., Tretyakov E.S., Korenkov V.V., Onykiy B.N., Artamonov A.A. // *Proceedings of the VIII International Conference "Distributed Computing and Grid-technologies in Science and Education" (GRID 2018)*, Sep. 2018, Vol. 2267, pp. 436-440.

9. Tretyakov E.S. Agent-based search in social networks / Artamonov A.A., Ionkina K.V., Kirichenko A.V., Lopatina E.O., Tretyakov E.S., Cherkasskiy A.I. // *International Journal of Civil Engineering and Technology (IJCET)*, Dec. 2018, Vol. 9, Issue 13, pp. 28-35.

10. Tretyakov E.S. Methodology of Analysis of Similar Objects with the Use of Modern Visualization Tools / Tretyakov E. S., Tukumbetova R. R. и Artamonov A. A. // *Advanced Technologies in Robotics and Intelligent Systems*, Jan. 2020, Vol. 80, pp. 113-119.

II. Публикации в журналах, включенных в перечень периодических изданий ВАК Российской Федерации

1. Третьяков Е.С. Анализ деятельности и информационного ресурса Агентства по перспективным оборонным научно-исследовательским разработкам США (DARPA)/ Артамонов А.А., Ананьева А.Г., Александрова Н.А., Кшняков Д.О., Третьяков Е.С. // *Наноинженерия*, № 8, 2015. – С. 43-48.

III. Публикации в других изданиях и материалы конференций

1. Третьяков Е.С. XXV Международный симпозиум по ядерной электронике и компьютерным технологиям (Symposium on Nuclear Electronics and Computing (NEC 2015) / Третьяков Е.С. //Будва, Октябрь 2015.
2. Третьяков Е.С. VII Международная конференция-школа «Распределённые вычисления и GRID-технологии в науке и образовании» / Третьяков Е.С. // Дубна, Декабрь 2016.
3. Третьяков Е.С. Международная конференция-школа «Проблемы обработки, анализа и управления Большими Данными в распределенной гетерогенной компьютерной среде для высокоинтенсивных областей науки и бизнес приложений» / Томск, Декабрь, 2016.
4. Третьяков Е.С. Международная конференция «Biologically Inspired Cognitive Architectures (BICA) for Young Scientists» / Прага, Август 2018.
5. Третьяков Е.С. VIII Международная конференция-школа «Распределённые вычисления и GRID-технологии в науке и образовании» / Дубна, Сентябрь 2018.
6. Третьяков Е.С. Исследование научно-технических разработок Китайской Народной Республики / Третьяков Е.С., Тукумбетова Р.Р. // Modern Science, № 4-3, 2019. – С. 123-126.

IV. Свидетельства о государственной регистрации баз данных

1. Третьяков Е.С., Петровский В. Н., Ионкина К.В., Артамонов А.А., Галин И.Ю., Оныкий Б.Н., Проничева Л.В., Данилова В.В., Кшняков Д.О., Суслина И.В., Черкасский А.И., Быковский Д.П., Интегральная база данных агентного поиска информации в сети Интернет по тематическому направлению Аддитивные технологии, Свидетельство о государственной регистрации базы данных 2017620244, февраль 27, 2017.
2. Третьяков Е.С., Оныкий Б.Н., Черкасский А.И., Проничева Л.В., Суслина И.В., Артамонов А.А., Ионкина К.В., Инкина В.А., Лопатина Е.О. База данных маркированных аудио-объектов, Свидетельство о государственной регистрации базы данных 2018621999, декабрь 11, 2018.

V. Свидетельства о государственной регистрации программ для ЭВМ

1. Третьяков Е.С., Ионкина К.В., Данилова В.В., Кшняков Д.О., Артамонов А.А., Оныкий Б.Н., Проничева Л.В., Суслина А.С., Толстая П.М., Программа автоматизированного квазиреферирования научно-технической информации, Свидетельство о государственной регистрации программы для ЭВМ 2017613798, март 31, 2017.
2. Третьяков Е.С., Оныкий Б.Н., Артамонов А.А., Ионкина К.В., Свистунов А.С., Лопатина Е.О., Зиновьева М.И., Черкасский А.И., Проничева Л.В., Суслина А.И., Иванченко А.М., Конвейерное определение тематической принадлежности научно-технической документации, Свидетельство о государственной регистрации программы для ЭВМ 2018665872, декабрь 11, 2018.