ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ. ВВЕДЕНИЕ

Лекция 1 – 2.

Литература

Интеллектуальная система, как черный ящик

Основные понятия

- **Определение 1.**
- Интеллектуальной называется система, способная целеустремленно в зависимости от состояния информационных входов, изменять не только параметры функционирования, но и сам способ своего поведения, причем способ поведения зависит не только от текущего состояния информационных входов, но также и от предыдущих состояний системы.

- Различают целенаправленные и целеустремленные системы. Примером системы первого типа может служить артиллерийский выстрел, второй — самонаводящаяся ракета.

- Любой живой организм — интеллектуальная система. Он обладает долговременной памятью и способностью к самообучению. Ребенок притронувшись к горячей плитке, уже не повторит ошибки. Щенок, впервые погнавшийся за кошкой, получит серьезный урок и вряд ли снова решит с ней поиграть. При следующей встрече он, скорее всего, убежит или покажет зубы, или проявит еще одну из тысяч возможных реакций.

- Технические же системы чаше всего не являются интеллектуальными, т. е. их реакция на одно и то же событие не может измениться кардинально. Система автоматического управления давлением газа в трубе может открывать и закрывать заслонку (управлять параметрами), но она не может принять решение совсем вывинтить заслонку из трубы. Если аварии газопровода предшествует изменение давления (например, сначала резкое повышение, а затем резкое понижение), то автоматическая система воспримет это как нормальную ситуация и попытается «отрегулировать» ее движением заслонки. Даже если после каждой аварии мы будет добавлять в систему управления новый блок, точно фиксирующий параметры предыдущей ситуации, ничего не изменится. Простое накопление данных не «обучит» систему. Дело в том, что щенок, получивший урок от кошки, запомнил не только параметры ситуации (длину когтей и скорость реакции), но и правила поведения (не подходи, не подставляй нос, если залаять — она убежит).

- **Определение 2.** Интеллектуальной называется система, моделирующая на компьютере мышление человека.

- 60-е годы, попытка смоделировать на компьютере мозг человека. Клетки мозга — нейроны — программно описывались специальными математическими методами. Компьютерная программа, таким образом, представляла как бы кусочек мозга человека. На вход программы подавались некоторые данные (на вход клетки мозга в живом организме подается электрический сигнал), на выходе снимались результаты, которые сверялись с эталоном. В зависимости от того, насколько полученные результаты отклонялись от эталона, в расчетные коэффициенты вносились изменения. В зависимости от количества циклов такого «обучения» результаты работы программы постепенно все более приближались к результатам работы очень маленького элемента мозга человека.

- **Определение 3.** Интеллектуальной называется система, позволяющая усилить интеллектуальную деятельность человека за счет ведения с ним осмысленного диалога.

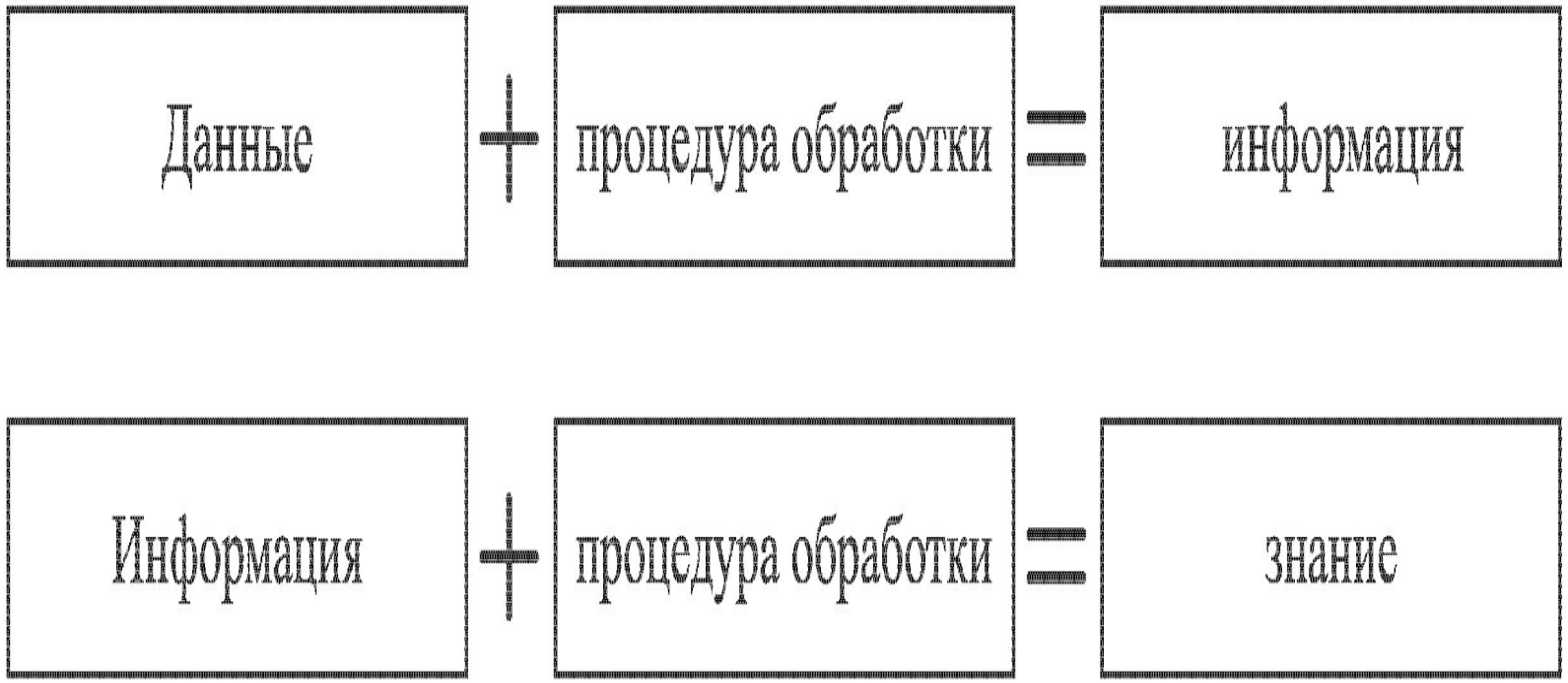
- Следует создавать узкоспециализированные интеллектуальные системы, которые не заменяют человека, но дополняют его. Человек имеет ряд уникальных способностей, но не свободен от недостатков. Не один человек не обладает реакцией кошки. Никто из нас не способен прочитать за минуту роман Л. Н. Толстого «Война и мир», редко кто из людей обладает энциклопедической памятью. Компьютер обладает энциклопедической памятью, компьютер совершает миллионы операций в секунду, компьютер реагирует мгновенно. Обратите внимание на новый акцент в постановке задачи создания ИИ. Если изначально выдвигалось требование к машине «мыслить», то теперь — «получать хорошие результаты». Другими словами, произошел переход от моделей, воспроизводящих процесс мышления человека или структуру головного мозга, к моделям, использующим какие-либо собственные принципы организации и методы обучения, но позволяющим получать результаты, «похожие на человека».

- Поясним примером: система автоматического наведения ракет обнаружила цель. Цель была обнаружена практически мгновенно, человек даже не успел ее заметить. Ракета была автоматически наведена на цель. Цели был послан запрос «свой-чужой». Цель появилась на пульте управления перед оператором, человек принял решение о поражении, выбрал тип оружия и нажал на кнопку «уничтожить». В случае полностью автоматического ведения цели существовала бы реальная опасность уничтожить свой самолет. В обратном случае, если бы наведением на цель, посылкой запроса занимался человек, могло быть упущено время.

- Ваша программа не станет интеллектуальной, если начнет заносить в свою базу данных все ситуации, с которыми она встречалась. Постоянно пополнять базы данных можно, и интеллектуальные программы это делают, однако это далеко не все.

- Рассмотрим пример. Человек смотрит на часы.
- Что он видит? Данные? Информацию? Знания?
- Проследим, насколько это возможно, действия человека. Итак, глаза смотрят на циферблат. В мозг поступает электрический сигнал, в мозге формируется изображение стрелок на циферблате. Далее, сознательно или подсознательно человек прикладывает некоторые умственные усилия, чтобы понять (по положению стрелок) сколько же сейчас времени (т. е. соотносит полученные данные с некоторой шкалой). Получив в итоге декларативную информацию, например 17:20, человек подключает внешние знания, например свое рабочее расписание, понимает что опаздывает (знания) и ускоряет шаг (меняет параметры своего поведения, т. е. данные).

Вывод на знаниях



- Характерная особенность знаний состоит в том, что они не содержатся в исходной системе. На циферблате часов не было написано «опаздываю». Слово «опаздываю» не содержалось и в расписании этого человека. Знания возникают в результате сопоставления информационных единиц, нахождения и разрешения противоречий между ними. Т. е. знания активны, их появление (или недостача) приводит к реализации некоторых действий.

- Для знаний характерны следующие свойства:
- -внутренняя интерпретируемость (каждая информационная единица должна иметь уникальное имя и однозначно определяться);
- -структурированность, т. е. между информационными единицами должны быть установлены отношения (например, «часть - целое», «род» — «вид» и др.); при этом возможна рекурсивность;
- знания образуют некоторое пространство, которое может оказаться как метрическим, так и не метрическим.

Современные теоретические проблемы ИИ.

- 1. Проблема представления знаний.
- 1.1. Разработка новых моделей представления для узкоспециализированных предметных областей.
- 1.2. Биомашины — машины, имеющие своей частью живые су-существа либо структурно подражающие человеку:
- 1.2.1. подражание моторике человека (походка, пластика, бег, прыжки, создание двуногих роботов);
- 1.2.2.создание инженерных моделей для различных областей по аналогии с системами человеческого организма.
- 1.3. Многокритериальное принятие решений.
- 1.4. Принятие решений на основе статистических моделей.
- 1.5. Координация работы нескольких роботов.
- 1.6. Проблемы совершенствования нейронных сетей.

Современные теоретические проблемы ИИ.

- 2. Проблема совершенствования компьютерной логики.
- 2.1. Разработка новых архитектур компьютеров (параллельные машины, исследования в области так называемой интегрированной памяти, децентрализованные машины, моделирование высокоскоростных электрических соединений).
- 2.2. Человекообразные роботы:
- 2.2.1. Гибкие и портативные члены роботов (головы, руки, тела и т. д.);
- 2.2.2. Распознавание роботами лиц, авторизованных для управления роботами;
- 2.2.3. Разработка механизмов роботов (человекоподобный палец с сенсорами, человекоподобная модель мускулатуры, говорение роботов, создание роботов для детей);
- 2.2.4. Численные методы для оптимизации вычислений.

Современные теоретические проблемы ИИ.

- 2.3. Методы доступа к информации:
- 2.3.1. Мультимедийные системы;
- 2.3.2. Эвристический анализ текстов;
- 2.3.3. Автоматическое извлечение знаний (ключевых слов) из текста;
- 2.3.4.Анализ авторского права на текст на основе образцов текста.

Современные теоретические проблемы ИИ.

- 2.4. Создание «интеллектуальных пространств»:
- 2.4.1. Интеллектуальные обучающие среды и оболочки;
- 2.4.2. Формирование запросов (к БД) на основе внимания, уделяемого пользователем различным элементам среды;
- 2.4.3. Менеджеры ресурсов для интеллектуальных пространств;
- 2.4.4. Автоматизация программирования и создания программной документации — интеллектуальная поддержка технологий разработки ПО (UML);
- 2.4.5. Многопользовательские среды.

Современные теоретические проблемы ИИ.

- 2.5. Машинное обучение:
- 2.5.1. Марковские процессы;
- 2.5.2. Машинное чтение и понимание текстов;
- 2.5.3. Восстановление утраченных элементов данных;
- 2.5.4. Очистка данных от шумов.
- 2.6. Медицинское зрение:
- 2.6.1. Автоматический анализ анатомических структур;
- 2.6.2. Чтение снимков (рентген и т. п.);
- 2.6.3. Машинная геометрия и пространственные сцены;
- 2.6.4. Восстановление изображений по их отражениям;
- 2.6.5. Сегментация изображений (например, простейшая разбивка на растр).

Современные теоретические проблемы ИИ.

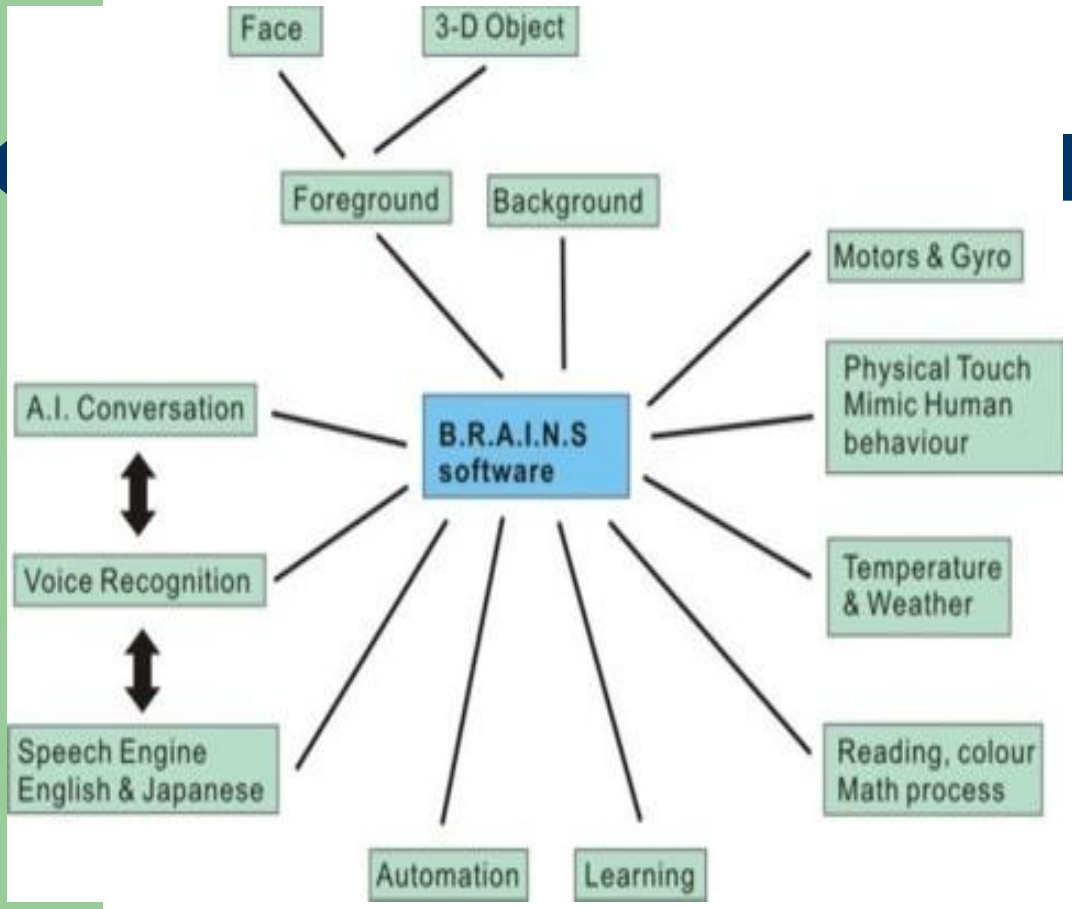
- 2.7. Мобильные роботы.
- 3. Проблема практического применения теоретических моделей.
- 4. Проблема совершенствования компьютерной лингвистики.
- 4.1. Разработка языка управления роботами на основе естественного языка.
- 4.2. Создание моделей естественного языка.
- 4.3. Понимание речи.
- 4.4. Разработка языков программирования, позволяющих повысить надежность разрабатываемого программного обеспечения (динамические языки, адаптивные системы, системы, способные «выживать»).

Asimo

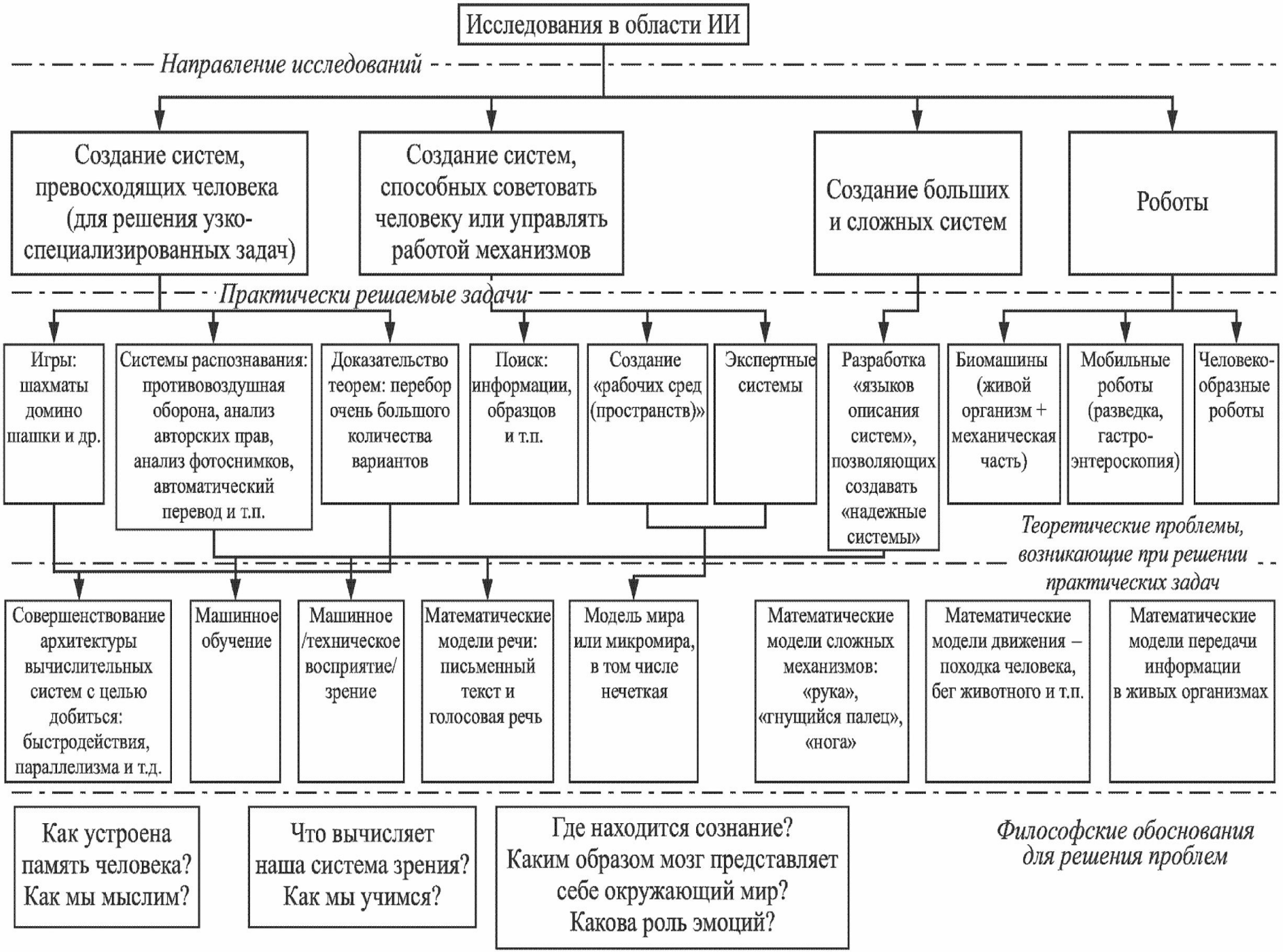
- Asimo (сокращение от Advanced Step in Innovative MObility) — робот-андроид. Создан корпорацией Хонда, в Центре Фундаментальных Технических Исследований Вако (Япония). Рост 130 см, масса 54 кг. Способен передвигаться со скоростью быстро идущего человека — до 6 км/ч.

Проект Aiko Спецификация:

- Height: 152cm
- Microprocessor: LS372 and C7 Board
- Micro Controller: 5-8
- 8Gig solid state HD with 4Gig internal memory
- Central Data 1000-1500Gig
- Power source: Mn-Polymer 7.2V and 12.0V
- Motors: Max 130kg.cm precise feedback is provided by the military grade, stainless steel potentiometer
- Motors speed: Max 0.12sec at 60 degree
- 23 (36) DOF
- 2x gyros
- 2cameras, 1x1ccd (1x3ccd)
- Control by Internal 32 bits OS


Face
3-D Object
Foreground
Background
Motors & Gyro
Physical Touch
Mimic Human
behaviour
A.I. Conversation
B.R.A.I.N.S
software
Voice Recognition
Temperature
& Weather
Speech Engine
English & Japanese
Reading, colour
Math process
Automation
Learning


EXIT


Исследования в области ИИ
Направление исследований
Создание систем, превосходящих человека (для решения узко-специализированных задач)
Создание систем, способных советовать человеку или управлять работой механизмов
Создание больших и сложных систем
Роботы
Практически решаемые задачи
Игры: шахматы домино шашки и др.
Системы распознавания: противовоздушная оборона, анализ авторских прав, анализ фотоснимков, автоматический перевод и т.п.
Доказательство теорем: перебор очень большого количества вариантов
Поиск: информации, образцов и т.п.
Создание «рабочих сред (пространств)»
Экспертные системы
Разработка «языков описания систем», позволяющих создавать «надежные системы»
Биомашины (живой организм + механическая часть)
Мобильные роботы (разведка, гастро-энтероскопия)
Человеко-образные роботы
Теоретические проблемы, возникающие при решении практических задач
Совершенствование архитектуры вычислительных систем с целью добиться: быстродействия, параллелизма и т.д.
Машинное обучение
Машинное /техническое восприятие/ зрение
Математические модели речи: письменный текст и голосовая речь
Модель мира или микромира, в том числе нечеткая
Математические модели сложных механизмов: «рука», «гнущийся палец», «нога»
Математические модели движения – походка человека, бег животного и т.п.
Математические модели передачи информации в живых организмах
Как устроена память человека? Как мы мыслим?
Что вычисляет наша система зрения? Как мы учимся?
Где находится сознание? Каким образом мозг представляет себе окружающий мир? Какова роль эмоций?
Философские обоснования для решения проблем

Главное препятствие на пути прогресса в области ИИ:

Дано:

1. Кубик А имеет черный цвет.
2. Кубики А и Б исходно расположены по отдельности друг от друга.

Верно ли утверждение, что *если кубик А положить на кубик Б, то кубик А по-прежнему останется черным*.

В рамках ситуационного исчисления для получения требуется наличие дополнительного утверждения: *«Кубик А не меняет своего цвета при такой операции»*

Это - тяжелая задача для систем ИИ, т.к. число подобных утверждений может быть весьма велико.

«Естественный закон инерции»: изменения в системе не происходят до тех пор пока они не оговариваются в системе заранее.

Другая сторона этой проблемы по М.Минскому: *каким образом из памяти робота можно удалить последствия некоторого действия, если ситуация изменилась и это действие отменяется*.

Или: *каким образом машинная программа определяет, какие из известных ему сведений робот должен пересмотреть, когда он намечает совершить то или иное действие?*

Применение методов ИИ для создания систем защиты информации

- Где применяются????

Где применяются???

- Компьютерная безопасность
  - Обнаружение внешних и внутренних вторжений
  - Моделирование и анализ поведения пользователей
- Электронный документооборот
  - анализ и фильтрация электронной почты и Web трафика
  - рубрикация и аннотирование электронных документов организации
- Технологические процессы и производство
  - выявление нештатных ситуаций
  - прогнозирование качества продукции

ИИ в компьютерной безопасности

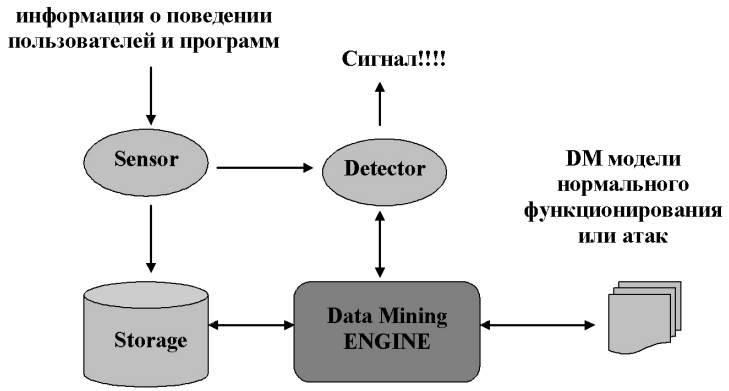
- Цели компьютерной безопасности: обеспечение конфиденциальности, целостности и доступности данных
- Вторжение – действия программы или пользователя, направленные на нарушение целей компьютерной безопасности
- Традиционные методы предотвращения вторжений (авторизация, разграничение прав доступа, криптозащита и т.д.) не справляются
- Необходимо выявление вторжений

Традиционные средства выявления вторжений

- Основные концепции:
  - Используют базы сигнатур известных атак
  - Источники информации: системные журналы и файлы, содержимое сетевого трафика и файлов.
- Недостатки:
  - Базы знаний формируются экспертами
  - Необходимо периодически обновлять
  - Существенная задержка во времени между появлением новой атаки и средств защиты от нее
  - Атаки постоянно видоизменяются
  - Есть методы «маскировки» атак

Методы ИИ в задачах выявления вторжений

- Основное предположение:
  - активность пользователей и программ можно полностью отследить и построить ее адекватную модель
- Особенности:
  - накопление исторической информации
  - модели нормального поведение или вторжения
  - эффективные методы анализа, которые проверяют текущую активность в системе на соответствие построенным моделям



Обнаружение нарушений

- Особенности:
  - Строится обобщенная модель атаки
  - Основано на методах классификации
  - Атакой считаются события или последовательности событий, соответствующие модели
- Основные проблемы:
  - «Обучение с учителем»: модель строится на примерах атак (необходимо их иметь и выделить из общей массы данных «вручную»)
  - Невозможно обнаруживать абсолютно новые или хорошо «замаскированные» атаки

Обнаружение аномалий

- Особенности :
  - Строится обобщенная модель нормальной активности пользователей или программ (профайл)
  - Основано на методах поиска исключений
  - Атакой считаются события или последовательности событий, несоответствующие модели
- Основные проблемы:
  - Предположения («Обучение без учителя»):
    1. обычные события отличаются от атак
    2. атак не больше p% от всех тренировочных данных, где p мало или равно 0 (обычно p неизвестно)
  - Высокий уровень ошибок второго рода (false positive)

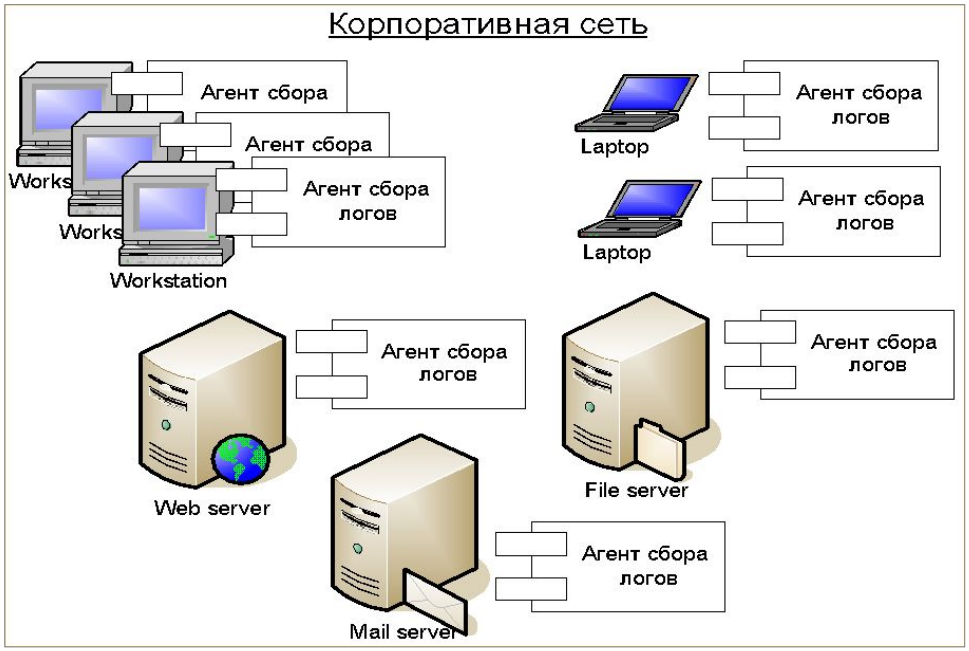
Разработанные и реализованные алгоритмы

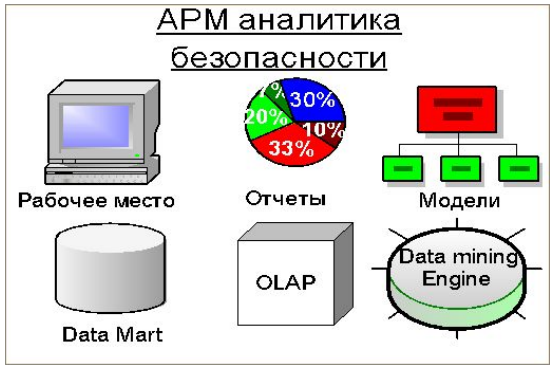
- Обнаружение аномалий:
  - Оценка степени «типичности» событий и их последовательностей - нечеткая кластеризация в бесконечномерном пространстве характеристик.
- Обнаружение нарушений:
  - Гибридный метод – Нечеткий SVM (Fuzzy Support Vector Machine) в сочетании с предыдущим методом
- «Описательные» модели поведения пользователей:
  - Вероятностная модель поведения пользователя на основе деревьев решений и отображения множества ситуаций (последовательностей событий) в пространство характеристик с помощью потенциальных функций
- Верификация:
  - На реальных данных и на эталонных тестовых наборах и др.

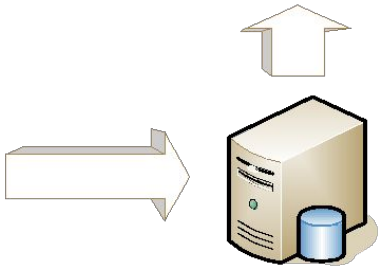
Система мониторинга и анализа поведения пользователей

- Функциональность:
  - Сбор и консолидация данных о работе пользователей
  - Статистический и интеллектуальный анализ
  - Построение и визуализация моделей поведения
  - Поиск аномалий в работе пользователей
- Области применения:
  - Выявление инсайдеров и предотвращение утечек информации
  - Поиск и анализ последствий вторжений
  - Система «раннего предупреждения»
  - Анализ производительности и целевого использования пользователями вычислительных средств организации

Архитектура системы мониторинга







Сервер консолидации и хранения событий безопасности

Особенности реализации

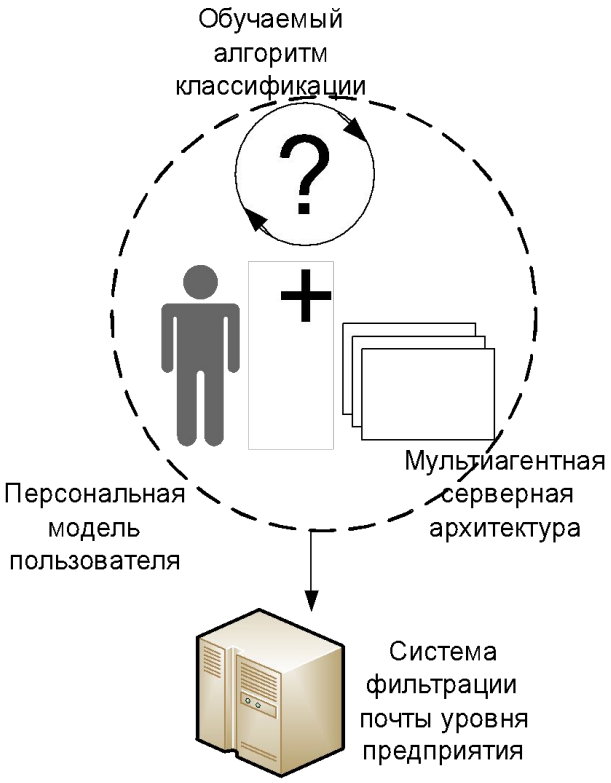
- Подсистема консолидации исходных данных:
  - Мульти-агентный подход
  - Нет ограничений на источники собираемых данных
  - Универсальный интерфейс для работы с модулями сбора данных
  - Специализированный формат представления собранных данных
  - Специализированное отказоустойчивое высоко производительное хранилище данных на файловой системе
  - Специальная предобработка данных
- Анализируемые факты:
  - Вход/выход в систему, запуск пользовательских и системных процессов, доступ к данным на любых носителях, активность пользователей в приложениях (клавиатура, мышь), входящий/исходящий сетевой трафик

Электронный документооборот

- Интеллектуальная система анализа и фильтрации электронной почты масштаба предприятия
- Система анализа и много-темной классификации Web трафика
- Интеллектуальная систему теневого копирования, рубрикации и аннотирования электронных документов организации

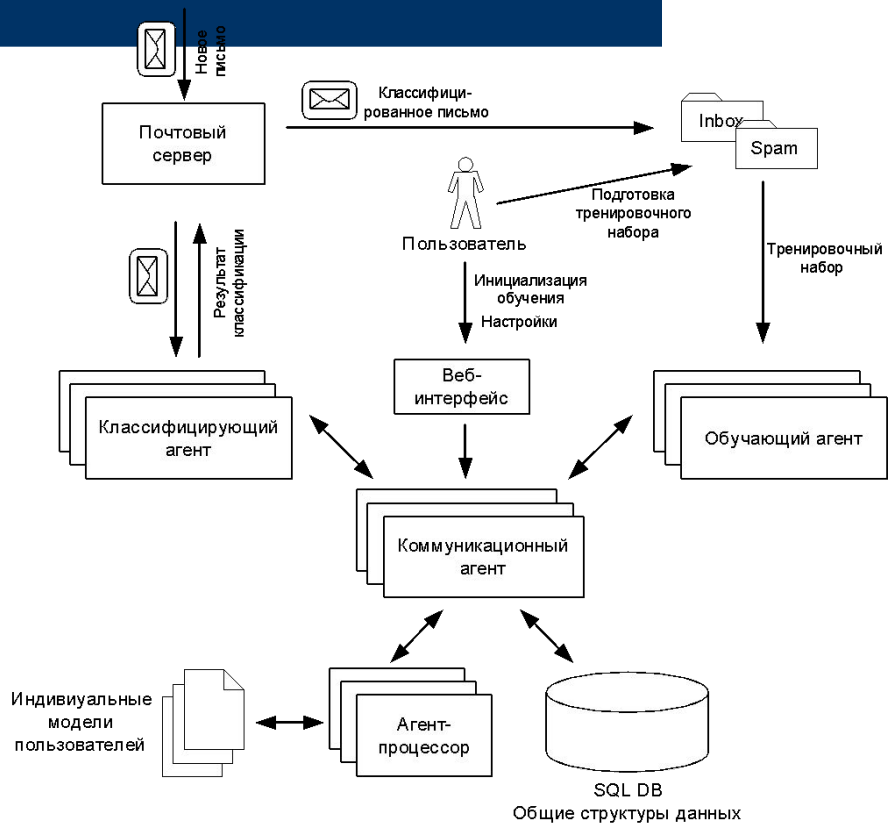
системы анализа и фильтрации электронной почты

- Алгоритм классификации (на SVM):
  - векторная форма представления письма
  - высокая точность
  - эффективность по скорости
  - персональная модель классификации почты
- Предобработка данных:
  - Снижение размерности исходного пространства (хи-квадрат и PCA)
  - Уменьшение размера тренировочного набора - кластеризация

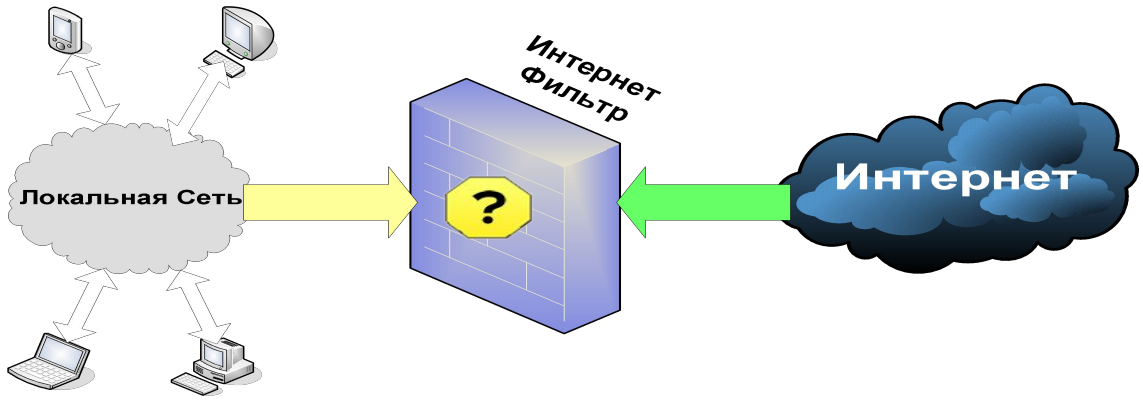


Архитектура системы фильтрации

- Особенности реализации:
  - Учет ресурсоемкости алгоритмов на этапе обучения
  - Распределение и баланс нагрузки
  - Классификация в режиме реального времени
  - Возможность масштабирования
  - Возможность интеграции с различными почтовыми системами



Цели создания систем анализа и фильтрации Интернет-трафика



- Блокирование доступа к нелегальной (экстремистской, антисоциальной, террористической и т.п.) информации
- Предотвращение использования Интернет-ресурсов в личных целях в рабочее и учебное время
- Предотвращение утечки конфиденциальной информации (анализ исходящего трафика)

Существующие системы фильтрации

- Традиционный подход («сигнатурные» методы):
  - Использование при анализе Интернет-трафика специализированных, формируемых экспертами, баз знаний, содержащих информацию об Интернет-ресурсах (URL, IP-адреса, ключевые слова)
- Основные недостатки:
  - Ориентированы на ресурсы со статическим содержанием («черные списки» адресов)
  - Возможны ошибки при определении тематики
  - Результаты зависят от качества и оперативности обновления баз знаний
  - Отсутствует анализа исходящего трафика (нет возможности предотвращения утечки конфиденциальной информации)

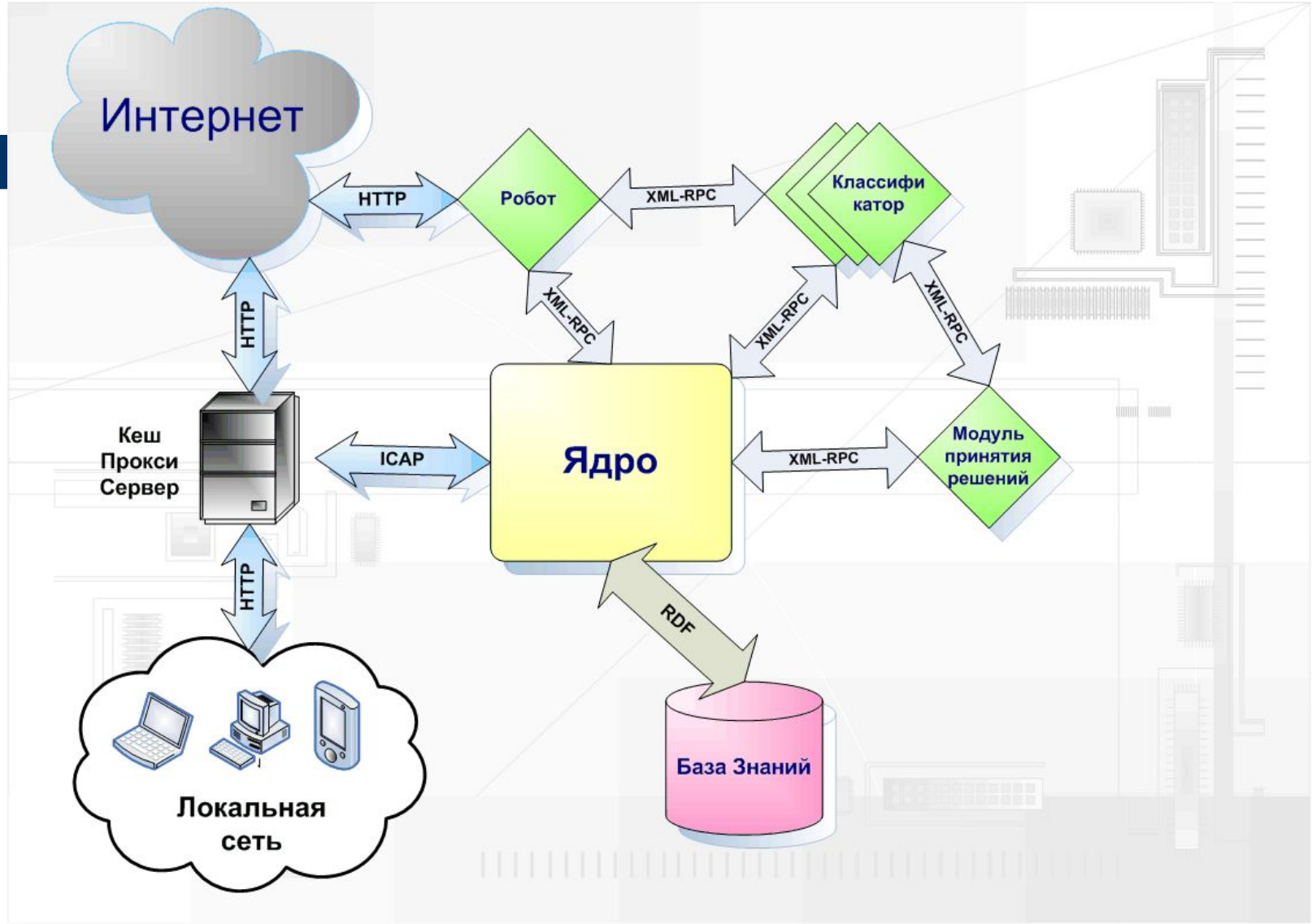
Анализ и фильтрация Интернет- трафика на основе методов ИАД

- Основная идея:
  - Классификация потока гипертекстовой информации в режиме реального времени с учетом содержания и структуры ссылок документов с использованием методов извлечения и применения знаний (алгоритмы машинного обучения и интеллектуального анализа данных).
- Функционирование:
  - Администратор формирует тренировочный набор с известными тематиками (примеры гипертексовых документов, либо список Интернет-ресурсов, содержимое которых затем откачивает робот);
  - На тренировочном наборе методами машинного обучения строится классификатор, который затем используется Интернет-фильтром в режиме реального времени для анализа содержимого трафика.
- На настоящий момент времени нет таких промышленных решений!

Преимущества

- Классификация в реальном времени статических и динамических интернет ресурсов;
- Точность выше, чем у «сигнатурных» методов;
- Автономность - независимость от внешних экспертов, поддержка собственной автоматически пополняемой базы знаний адресов;
- Адаптируемость - возможность уточнения классификации при поступлении новых примеров;
- Расширяемость - возможность добавлять новые категории и гибко настраивать политики фильтрации.

Архитектура системы



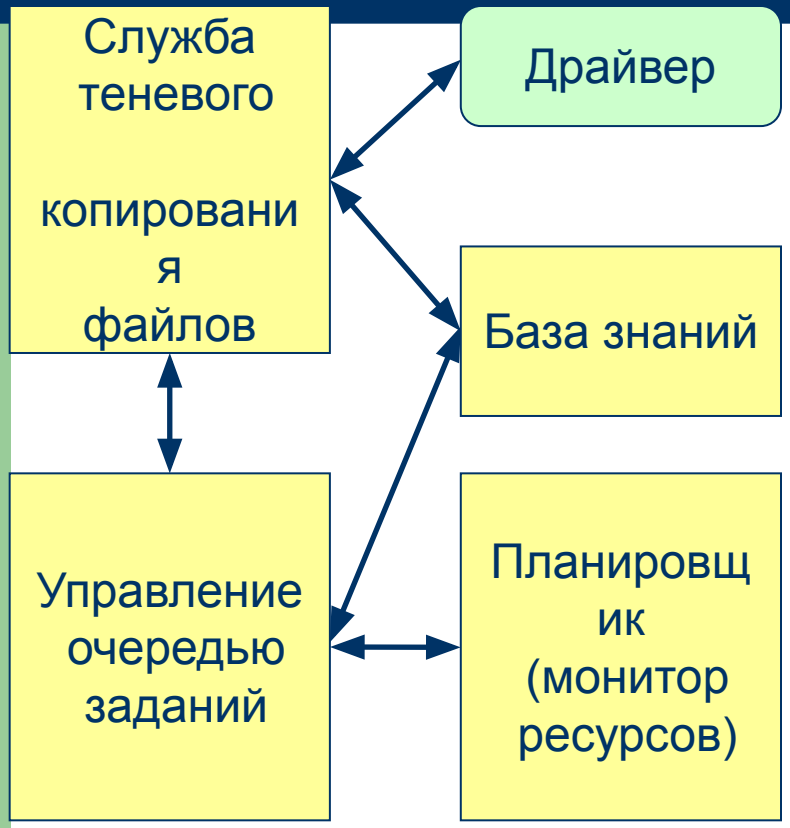
Основные результаты

- Реализация системы:
  - Формализованы требования и сценарии взаимодействия
  - Спроектированы и реализованы базовые компоненты, их функционал, интерфейсы, алгоритмы работы
  - Разработана онтология представления информации об интернет ресурсах и алгоритмы работы с базой знаний
- Разработан новый алгоритм много-темной классификации:
  - на основе модифицированного для существенно пересекающихся классов метода «попарных сравнений» с помощью набора бинарных классификаторов и отсечением нерелевантных классов
- Предложена расширенная векторная модель представления гипертекстовых документов:
  - включает базовые текстовые и нетекстовые признаки, составные признаки (сгруппированные базовые) определяются с помощью метода поиска частых эпизодов
  - новый метод учета гиперссылок (не требует загрузки содержимого «окружения»)

Интеллектуальная система анализа и мониторинга электронного документооборота организации

- Основная задача системы:
  - Перехват, «теневое копирование» и автоматизированное формирование «базы знаний» электронных документов организации
- Возможности системы :
  - журналируется история работы пользователей с документами и история изменений документов
  - для каждой версии документа автоматически определяется тематика, множество похожих документов (кластер), строится и сохраняется аннотация – набор ключевых фрагментов текста документа
  - выявление ключевых характеристик – алгоритмы SVD, ICA и др.
  - администратор может выполнить поиск и классификацию документов по содержимому и по аннотациям

Архитектура



Драйвер ФС: определяет с какими файлами работал пользователь;

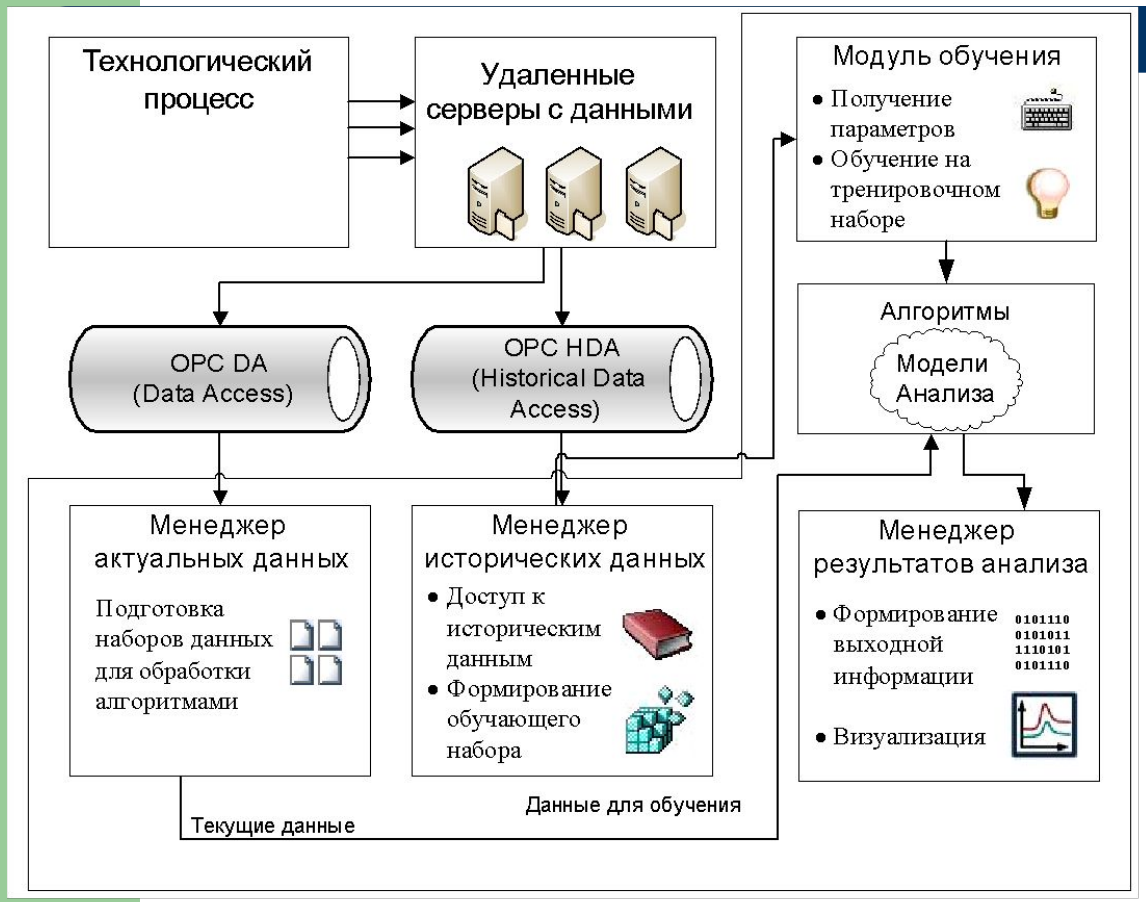
Служба теневого копирования: определяет как сильно изменился файл, при необходимости делает резервную копию, передает файл на обработку;

База знаний: хранение резервных копий файлов их аннотаций, служебной информации о кластерах и моделей аннотирования;

Управление очередью заданий: хранит очередь заданий на обработку, при освобождении ресурсов ВС выполняет задания из очереди;

Монитор ресурсов: анализирует загруженность ВС, разрешает выполнять задания из очереди;

Архитектура ИАД системы анализа поведения технологических процессов

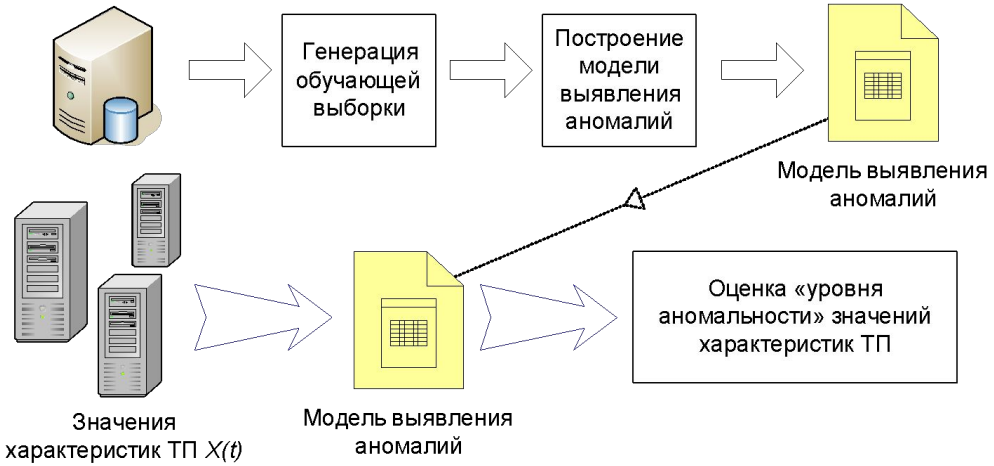


Особенности реализации:

- выявление аномалий в характеристик ТП
- функционирование в промышленной среде
- работа в режиме мягкого реального времени
- расширяемость по набору методов анализа

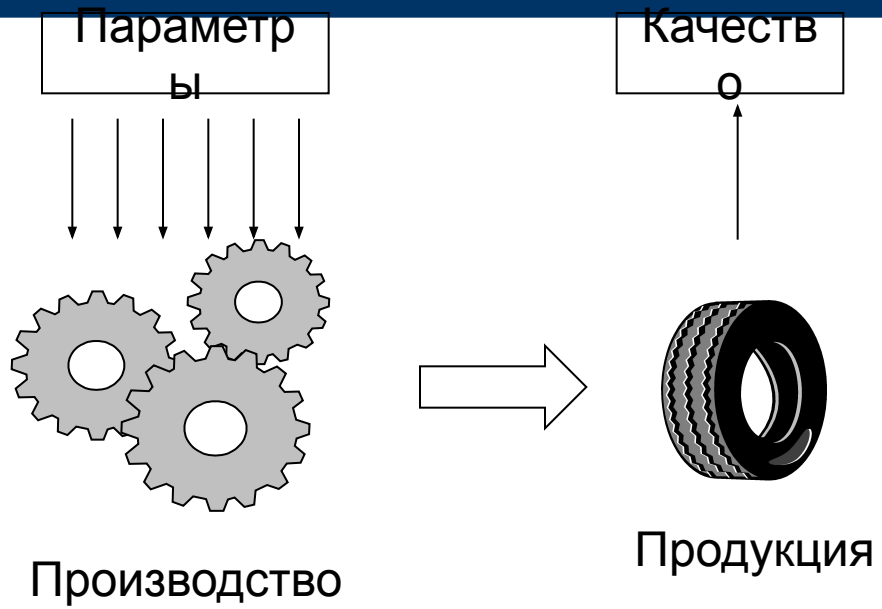
Выявление нештатных ситуаций

- построение модели поведения ТП (на этапе обучения)
- оценка отклонения текущего состояния ТП от модельного
- используются методы анализа временных рядов и последовательностей:
  - Класса «Гусеница» (Singular Spectrum Analysis)
  - Методы авторегрессии на основе SVR
  - Скрытые модели Маркова
  - и др.



Анализ и прогнозирование качества ТП

Какие параметры производственного процесса влияют на качество продукции?



$$Quality = F(X_1, \ldots X_n),$$

где X_i — i-ая характеристика производственного процесса

Результат

- Разработаны алгоритмы:
  - на основе нечетких деревьев решений
  - с поддержкой эволюционных методов оптимизации нечетких переменных и структуры правил
- Реализована экспериментальная программная система:
  - строит модели зависимости качества продукции от характеристик производственного процесса, представимую в виде системы нечетких правил «если … то … иначе»;
  - прогнозирование ожидаемого качества изделия по характеристикам производственного процесса производится с достаточной точностью;
  - позволяет упорядочить характеристики технологического процесса по степени влияния на качество.