



Методы интеллектуального анализа данных и некоторые их приложения

д.ф.-м.н. И.В.Машечкин (mash@cs.msu.su),
к.ф.-м.н. М.И. Петровский (michael@cs.msu.su)

лаборатория «Технологий программирования»
ВМиК МГУ им. М.В. Ломоносова

(<http://synthesis.ipi.ac.ru/sigmod/seminar/s20090226>)

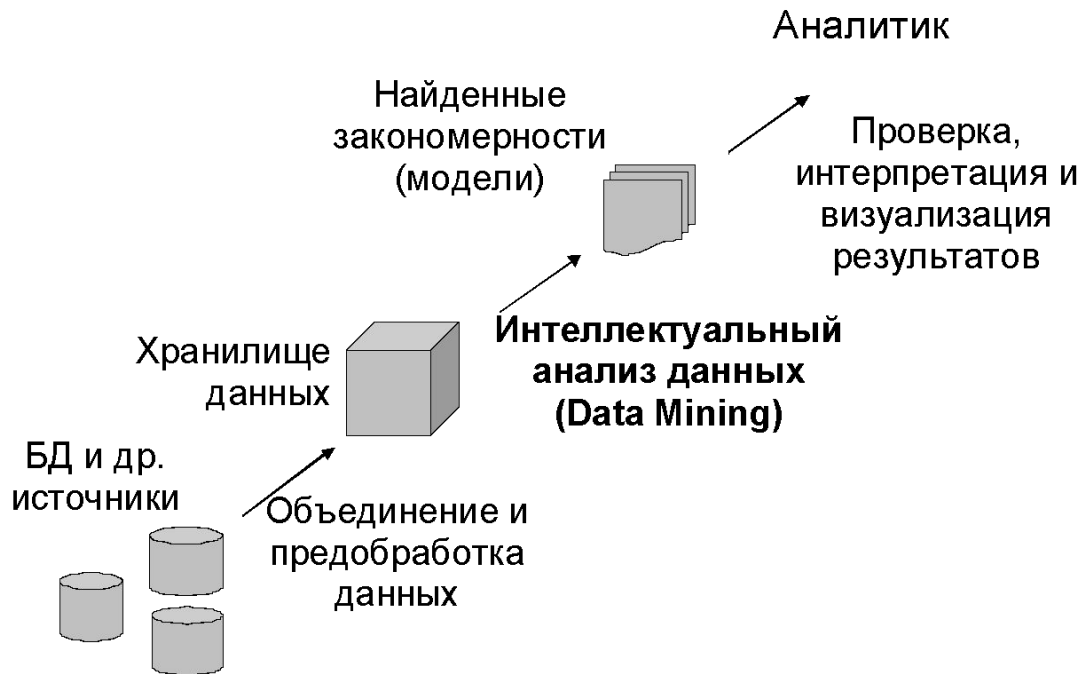
Эволюция технологий хранения и обработки данных

- ... — 1960-е:
 - Файлы и файловые архивы
- 1960-е:
 - Первые СУБД, иерархические, сетевые и т.д.
- 1970-е:
 - Реляционная модель данных, реляционные СУБД
- 1980-е:
 - «Продвинутые» СУБД (объектно-реляционные и объектные, «расширенные» реляционные, дедуктивные и д.р.)
 - «Специализированные» СУБД (гео-, научные, инженерные и д.р.)
- 1990-е —:
 - Мультимедийные БД, WWW, хранилища,
 - витрины данных, OLAP, Data Mining

Актуальность и необходимость интеллектуального анализа данных

- Проблема больших объемов («Data explosion»):
 - Средства автоматического сбора данных, повсеместное внедрение СУБД, электронный документооборот, WWW, мультимедийные архивы и т.д. Все ведет к росту объемов и усложнению структуры хранимой информации.
- Традиционные средства не справляются:
 - Информационный поиск и стат. анализ не везде помогают – много данных, сложная структура и нужно знать точно, что искать.
 - Вывод: много данных, но мало информации для аналитика.
- Необходимо:
 - Разработка программных средств автоматизированного анализа данных большого объема и сложной структуры.

Интеллектуальный анализ данных (Data Mining)



Системы *интеллектуального анализа данных* (ИАД) – класс программных систем поддержки принятия решений, задачей которых является поиск скрытых, ранее неизвестных, содержательных и потенциально полезных закономерностей в больших объемах разнородных, сложно структурированных данных.

Han J., Kamber M. Data Mining: Concepts and Techniques // Morgan Kaufmann, 2000

Процесс ИАД (1)

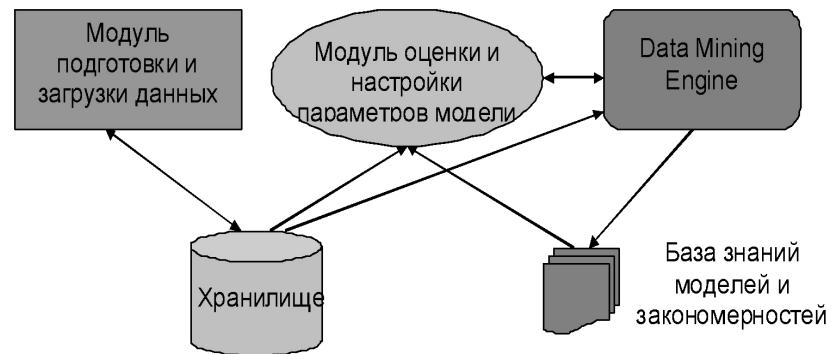
- Анализ предметной области:
 - выявление и формулировка необходимых априорных знаний о предметной области, целей анализа, задач приложения, сценариев использования
- Формирование и подготовка данных для анализа:
 - поиск (или выбор) «сырых» данных, возможно реализация подсистемы сбора (консолидации)
 - предобработка данных (нормализация, дискретизация, обработка пропущенных значений, удаление артефактов, проверка консистентности)
 - уменьшение размерности, выбор значимых характеристик, расчет интегральных показателей и инвариантов
- Определение типа решаемой задачи анализа:
 - классификация, прогнозирование, кластеризация, поиск исключений, ассоциативный анализ и т.д.

Процесс ИАД (2)

- Выбор (или разработка) алгоритма анализа:
 - определение ограничений и требований к алгоритму по точности, размеру, интерпретируемости, скорости построения и применения получаемых моделей, по типу исходных данных
- Собственно «Data mining»:
 - применение выбранного алгоритма анализа для поиска закономерностей выбранного типа и построение моделей
- Проверка моделей и представление результатов анализа:
 - визуализация, преобразование, удаление избыточности, оценка точности, достоверности моделей и т.д.
- Применение построенных моделей:
 - Descriptive data mining - информирование аналитика, «описательные» модели, основная цель – визуализация
 - Predictive data mining – прогнозирование неизвестных значений или характеристик в «новых» данных с помощью построенных моделей , основная цель – прогноз

Программные системы ИАД

- Типовая архитектура:



- Классификация систем ИАД:

- По типу анализируемых данных
- По типу решаемых задач
- По методам анализа и классам алгоритмов
- По области применения

Типы исходных данных (1)

- Транзакционные базы данных и репозитории «событий»
 - Объекты анализа – «события» различной структуры с числовыми и категориальными атрибутами, и с временной меткой
- Реляционные и объектные СУБД
 - Объекты анализа сложным образом взаимосвязаны (заданно ER-схемой), имеют разнотипные атрибутами, наследование (расширение)
- Многомерные OLAP-хранилища
 - Объекты анализа – срезы многомерно OLAP куба, т.е. набор числовых мер, при фиксированных значениях измерений
- Временные ряды и числовые данные большого объема
 - Обработка результатов наблюдений, научных экспериментов, характеристик технологических процессов

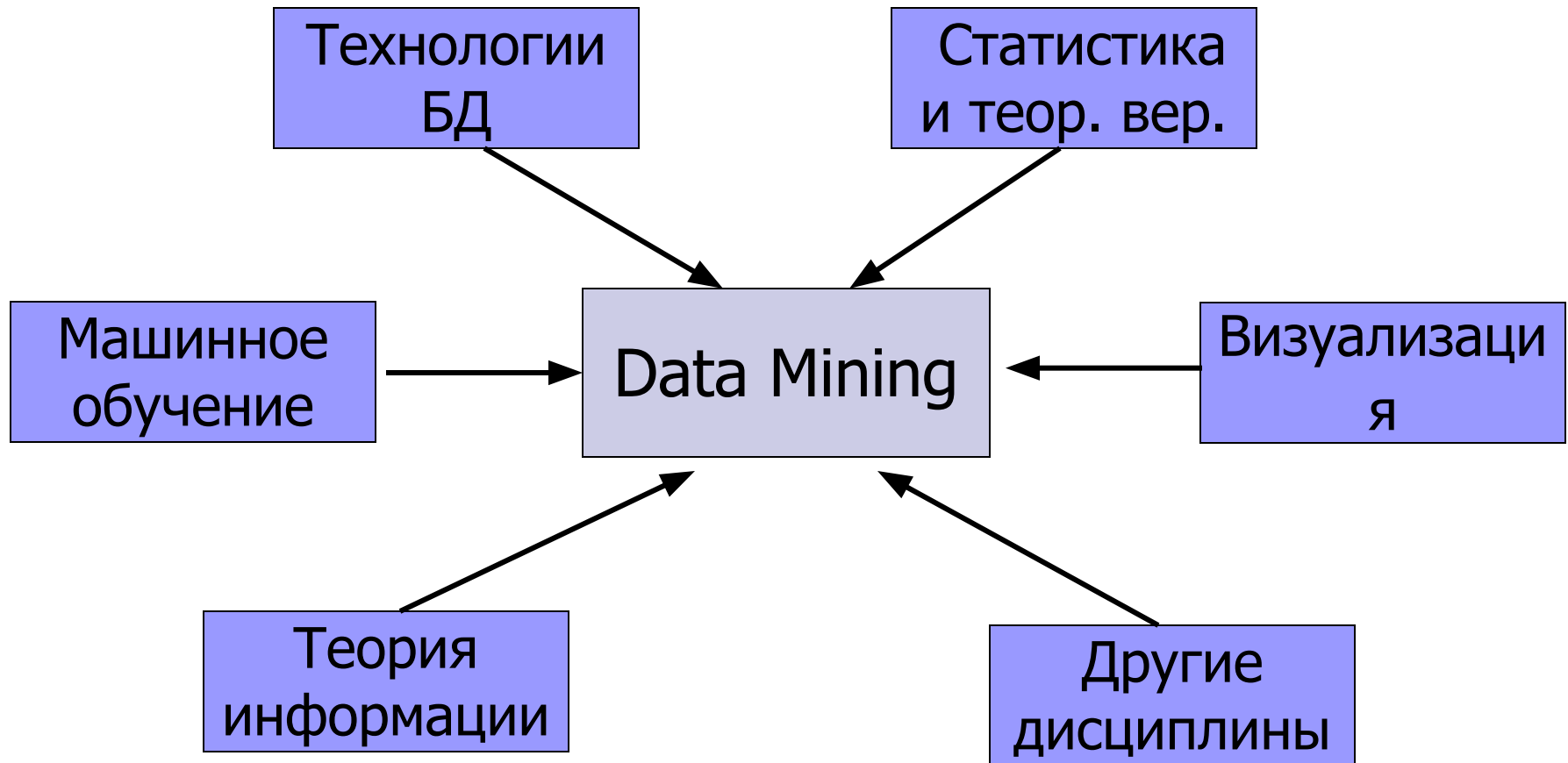
Типы исходных данных (2)

- Географические и пространственные данные
 - Привязка к пространственным координатам, учет географии объектов при анализе (например при определении меры сходства или расстояния) , учет перемещения в пространстве (moving objects)
- Символьные последовательности
 - ДНК цепочки, машинные коды, трассы выполнения процессов, тексты программ на ЯП
- Электронные тексты на естественном языке
 - анализ содержимого документов, проблема представления, морфология
- Гипертекстовые данные и WWW
 - структурированный текст на естественном языке, учет гиперссылок и нетекстового содержания
- Мультимедия
 - Звук, видео, изображения

Задачи ИАД = типы выявляемых закономерностей

- Классификация («Обучение с учителем»)
 - Отнесение объектов к заранее определенным категориям
- Прогнозирование («Обучение с учителем»)
 - На основании известных значений атрибутов анализируемого объекта определяются значения неизвестных атрибутов
- Ассоциации («Обучение без учителя»)
 - Выявление зависимостей между атрибутами
- Кластеризация («Обучение без учителя»)
 - Выделение компактных подгрупп «похожих» объектов
- Дискриминантный анализ («Обучение без учителя»)
 - Выявление атрибутов который «различают» (дискриминируют) две или более возникающие совокупности (группы)
- Выявление исключений («Обучение с и без учителя»)
 - Поиск объектов, которые своими характеристиками значительно отличаются от остальных

Методы анализа



Область применения систем ИАД

■ Системы ИАД «общего назначения»

- По сути включают framework, библиотеку алгоритмов анализа и набор программных средств для реализации ИАД процесса для широкого класса входных данных и прикладных задач
- Примеры DataMiner, MS Analysis Services, Oracle BI, PolyAnalyst

■ Специализированные системы ИАД

- Набор решаемых задач и алгоритмов решения, а также средств подготовки данных и визуализации результата ориентирован на конкретную предметную область
- ИАД процесс максимально «автоматизирован», но конечным потребителем информации все равно является эксперт-аналитик
- Области применения: маркетинг, анализ финансовых рисков, здравоохранение, страхование, кредитование, телекоммуникации, компьютерная безопасность, мониторинг оборудования и технологических процессов, антитерроризм, интернет и т.д.

Отличия ИАД систем (1)

- Наличие «обучения»
 - база знаний формируются на основе анализируемых данных, а не экспертных знаний (в отличии от традиционных экспертных систем и систем информационного поиска)
 - структура модели и искомые зависимости заранее не известны (в отличии от статистических пакетов, ориентированных на расчет статистик, проверку гипотез и оценку параметров распределений)

Отличия ИАД систем (2)

- Наличие большого объема данных сложной структуры
 - зачастую скорость работы алгоритмов в ИАД важнее небольших отклонений по точности (“quick and dirty solution”)
 - большинство алгоритмов работают с исходными данными в виде числовой матрицы признаков, сложная структура реальных объектов в ИАД, приводит к необходимости решать задачу построения пространства характеристик и отображения в него свойств исходных объектов
 - перечисленные особенности отличают ИАД системы от традиционных систем машинного обучения, в которых как правило решается обратная задача – построение достоверной модели в условиях малой обучающей выборки

Отличия ИАД систем (3)

- Наличие человека - аналитика как окончного потребителя результатов работы ИАД системы
 - в сценарии работы любой системы ИАД всегда присутствует аналитик, даже если полученная в результате модель далее используется для автоматической классификации
 - аналитик формирует тренировочные наборы, производит настройку алгоритмов, обучение и дообучение, анализирует полученные модели и принимает решения об их дальнейшем использовании
 - таким образом, системы автоматические классификации, кластеризации и распознавания образов, даже использующие возможность дообучения, не являются системами ИАД

ИАД в проектах лаборатории «Технологий Программирования»

- 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 нестандартных ситуаций
 - прогнозирование качества продукции
- Системы поддержки принятия решений
 - использование ИАД в ПО ситуационных центров

ИАД 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 базах 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

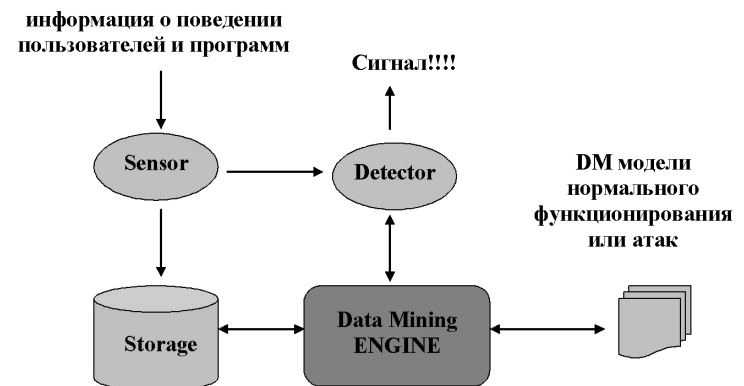
Методы ИАД в задачах выявления вторжений

■ Основное предположение:

- активность пользователей и программ можно полностью отследить и построить ее адекватную модель

■ Особенности:

- накопление исторической информации
- модели нормального поведения 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 выдeltь 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 DARPA и др.

Система мониторинга и анализа поведения пользователей

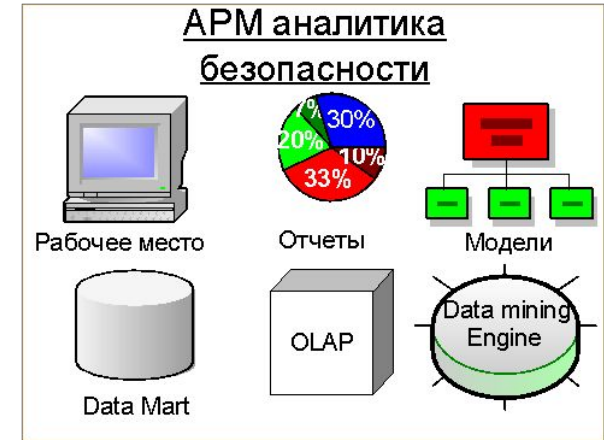
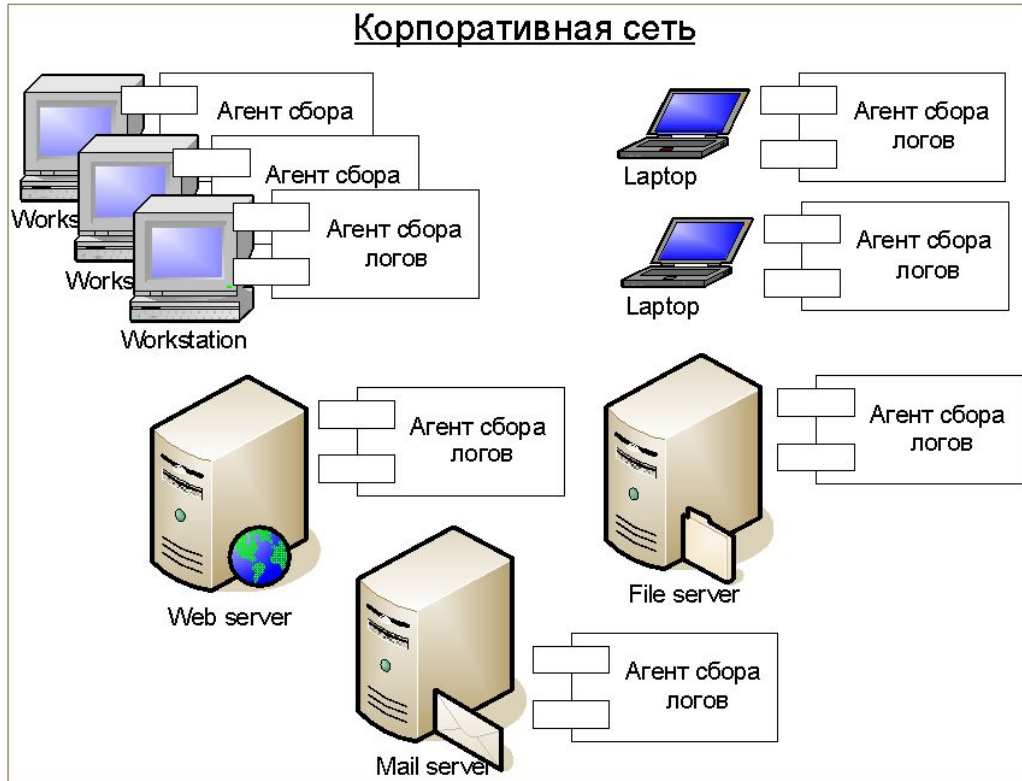
■ Функциональность:

- 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



Особенности реализации и результаты

- 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
- Опытная эксплуатация:
 - В ряде коммерческих и государственных организаций прошло опытное внедрение

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

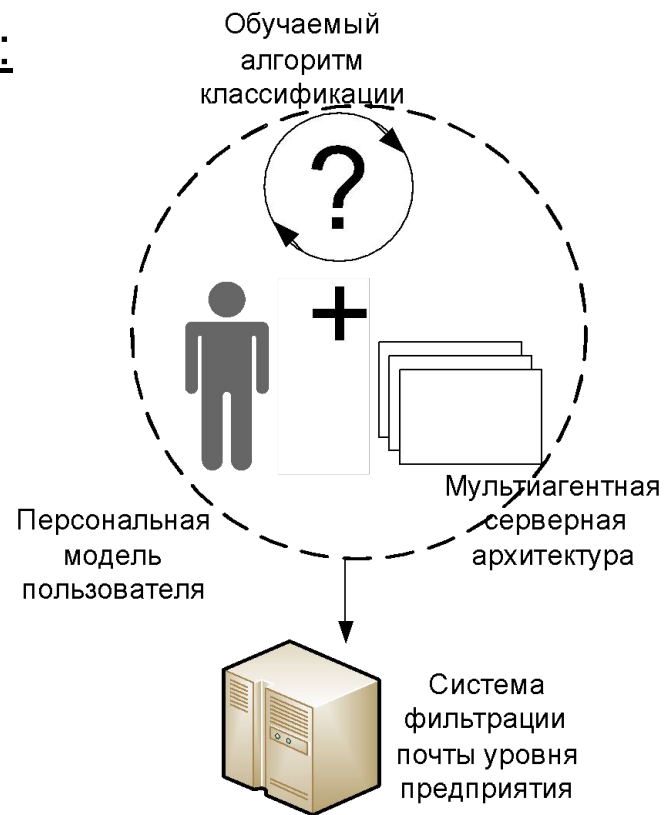
ИАД для системы анализа и фильтрации электронной почты

■ Алгоритм классификации (на SVM):

- векторная форма представления письма
- высокая точность
- эффективность по скорости
- персональная модель классификации почты

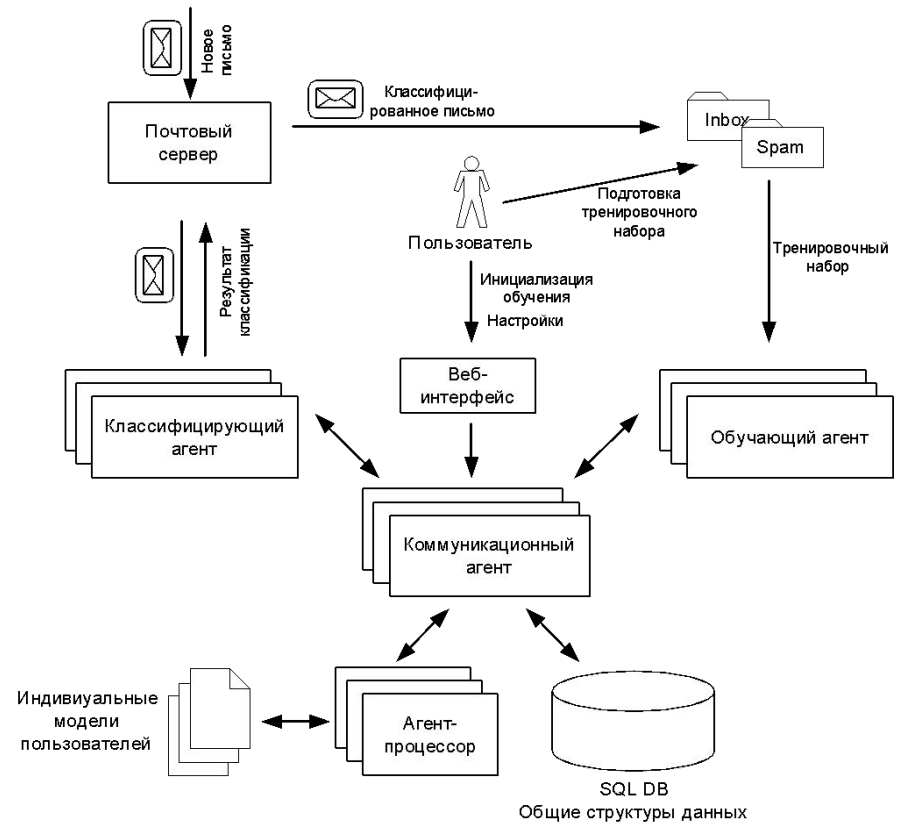
■ Предобработка данных:

- Снижение размерности исходного пространства (хи-квадрат и PCA)
- Уменьшение размера тренировочного набора - кластеризация



Архитектура системы фильтрации

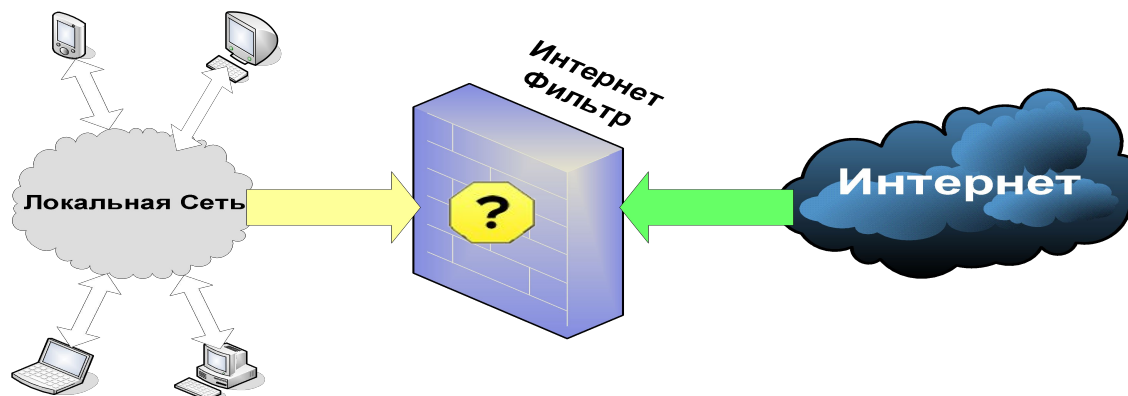
- Особенности реализации:
 - Учет ресурсоемкости алгоритмов на этапе обучения
 - Распределение и баланс нагрузки
 - Классификация в режиме реального времени
 - Возможность масштабирования
 - Возможность интеграции с различными почтовыми системами



Результаты экспериментальной реализации и апробации

- Почтовый сервер лаборатории «Технологий программирования»
 - эксплуатация с весны 2004
 - около 1 тыс. писем в день (после RBL)
 - из них > 70% спам
 - уровень обнаружения более 95%
 - уровень ложно-положительных ошибок ~ 0.1%
- Почтовый сервер факультета ВМиК, МГУ
 - эксплуатация с осени 2004

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)

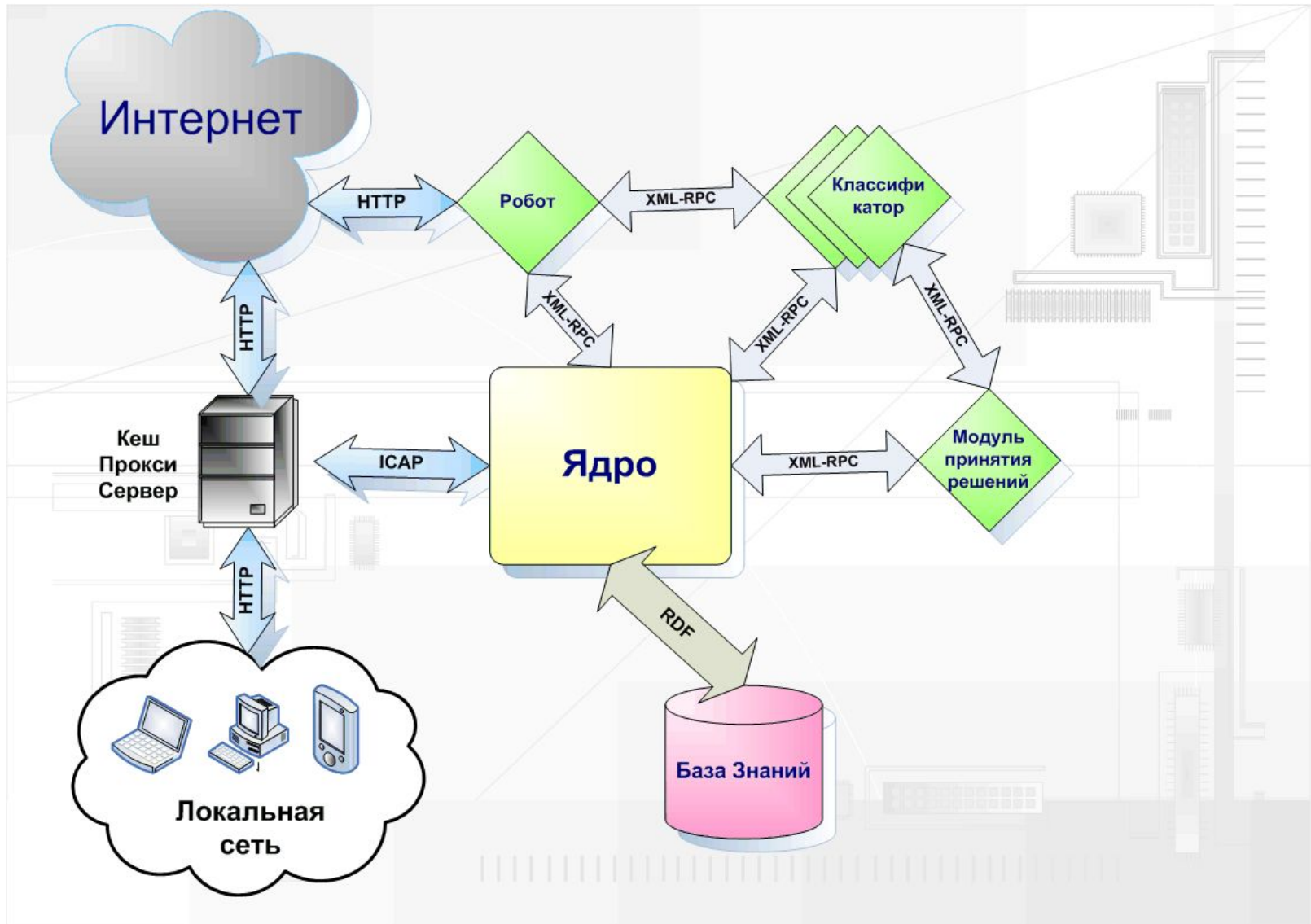
Анализ и фильтрация Интернет-трафика на основе методов ИАД

- 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 гипертекстовых 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.

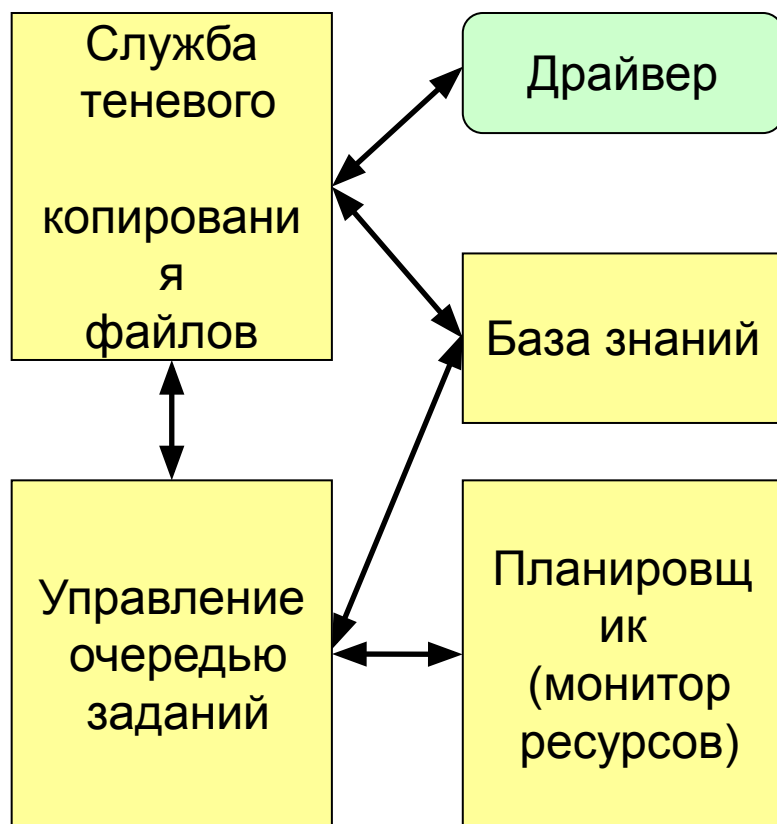
Архитектура системы



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



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;

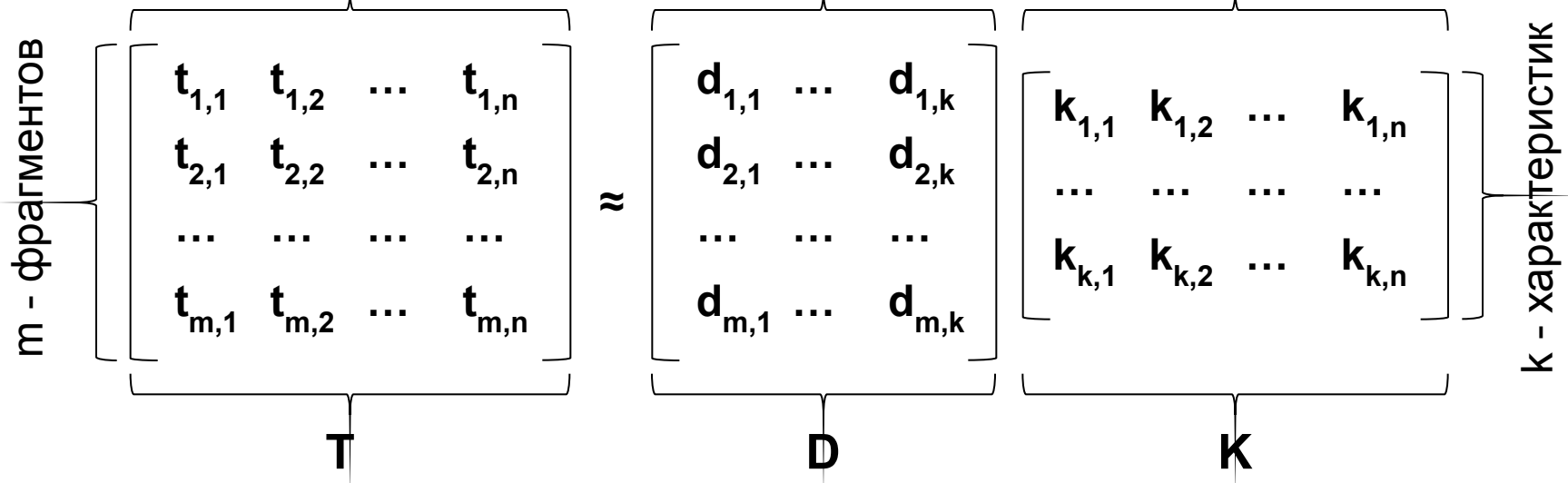
База знаний: хранение резервных копий файлов их аннотаций, служебной информации о кластерах и моделей аннотирования;

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;

Модель аннотирования

- *Кластер документов* – набор «схожих» документов;
- Каждый документ кластера разбивается на *фрагменты текста*;
- Каждый *фрагмент текста* преобразуется в частотный вектор «термов» (слов);



T – исходная матрица фрагментов;

D – матрица фрагментов в пространстве «ключевых» характеристик;

K – матрица перехода, K^{-1} – модель аннотирования (строится для каждого кластера);

Алгоритмы поиска ключевых характеристик

- *Латентно-семантический анализ (LSA - Latent Semantic Analysis):* основан на использовании разложения исходной матрицы по сингулярным значениям (SVD - разложение)
- *Анализ независимых компонент (ICA - Independent component analysis):* поиск линейных комбинаций наблюдаемых переменных ведется чтобы получить независимые случайные величины, распределение которых максимально далеко от нормального
- *Выделение частых эпизодов термов (Apriori):*
Для выделенных фрагментов документов, строится список характерных частых эпизодов термов.

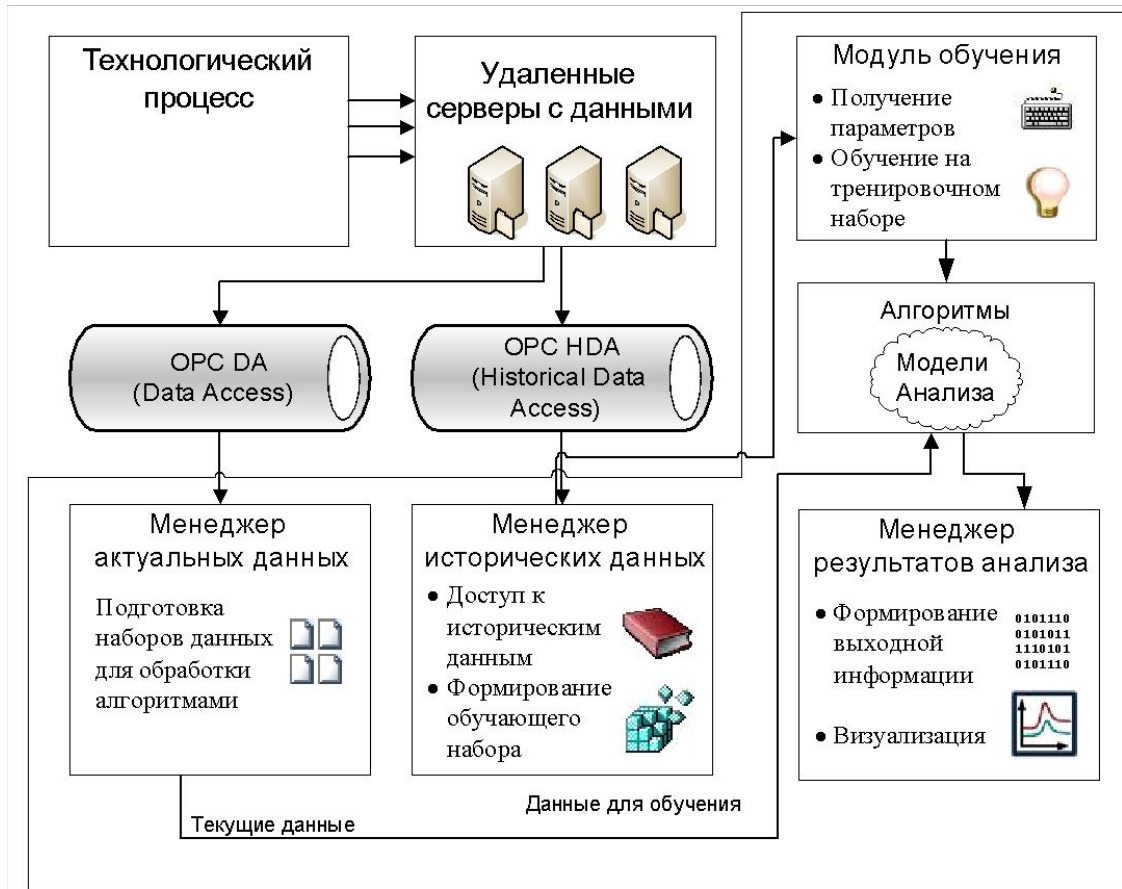
Аннотирование новых документов

- Для каждого *документа* определяется ближайший кластер и соответствующая ему модель аннотирования;
- Документ разбивается на *фрагменты текста*;
- Каждый *фрагмент текста* преобразуется в частотный вектор «термов» t_i ;
- Для каждого частотного вектора «термов» t_i строится его *проекция* на пространство ключевых характеристик p_i с помощью модели аннотирования K^{-1} :

$$p_i = t_i * K^{-1}; \quad K_{V_i} * K^{-1} = \sum_{j=1}^k E_j |p_{i,j}|;$$

- *Вес фрагмента* определяется как:
- *Аннотация* – n фрагментов с максимальным весом;

Архитектура ИАД системы анализа поведения технологических процессов

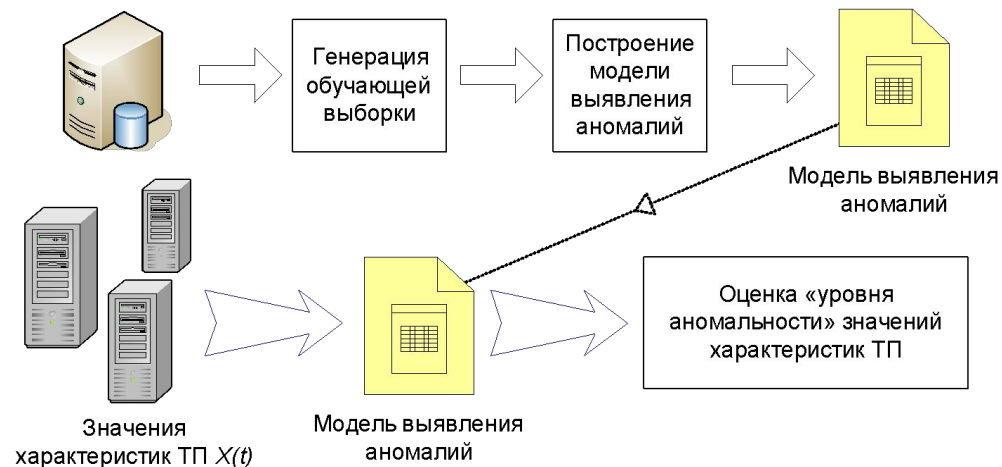


Особенности реализации:

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

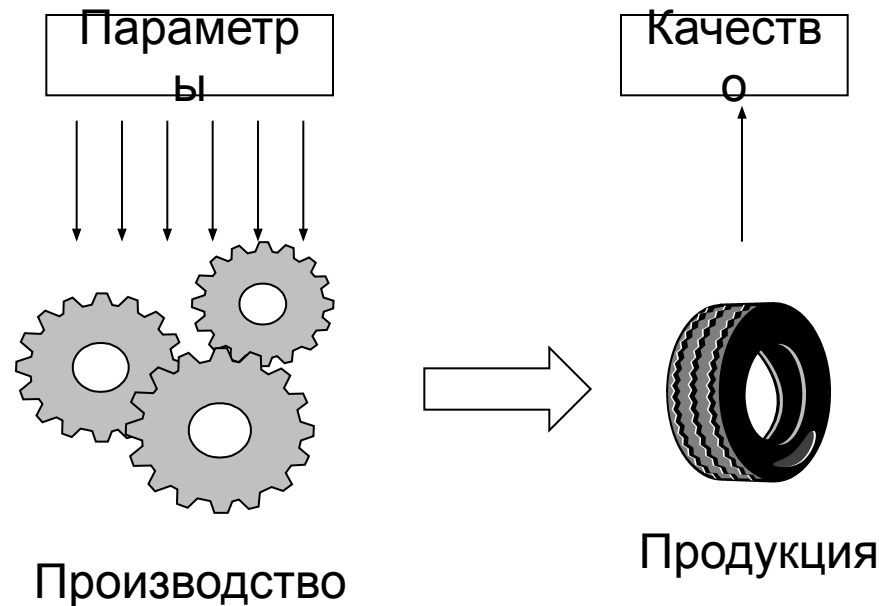
Выявление нештатных ситуаций

- 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?



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

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.

Ситуационный центр

- Основная задача СЦ — строить наглядные образы ситуаций, возникающих в предметной области, на основе которых оперативный состав принимает управляющие решения.
 - в СЦ обязательно входит оперативный состав (коллектив потребителей наглядной информации), решающий некоторую совокупность задач, требующих принятия решений;
 - в СЦ создаются информационные модели и картины весьма сложных, комплексных, динамических ситуаций реального мира для представления оперативному составу.
- Определение СЦ: это совокупность программно-технических средств, научно-математических методов и инженерных решений для автоматизации процессов отображения, моделирования, анализа ситуаций и управления.

Место ИАД в процессе поддержки принятия решений в СЦ

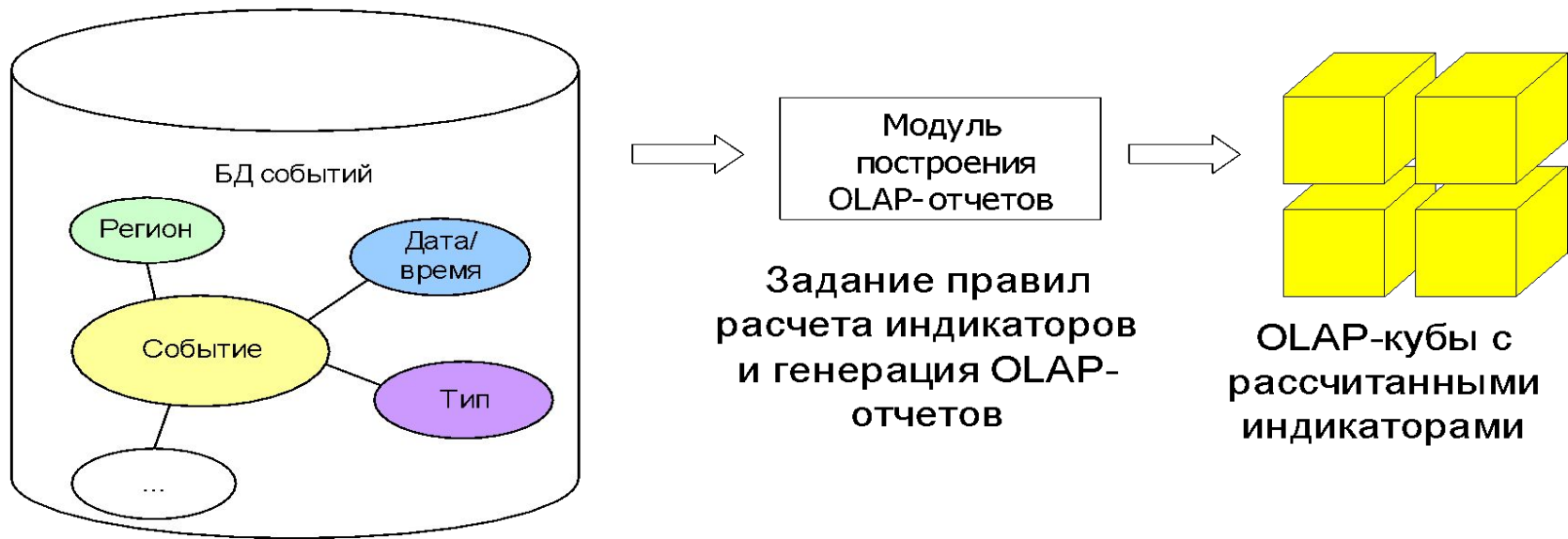
Задачи:

- Расчет индикаторов на основе данных предметной области
- Определение тенденций и прогнозирование значений индикаторов
- Выявление аномалий в значениях индикаторов

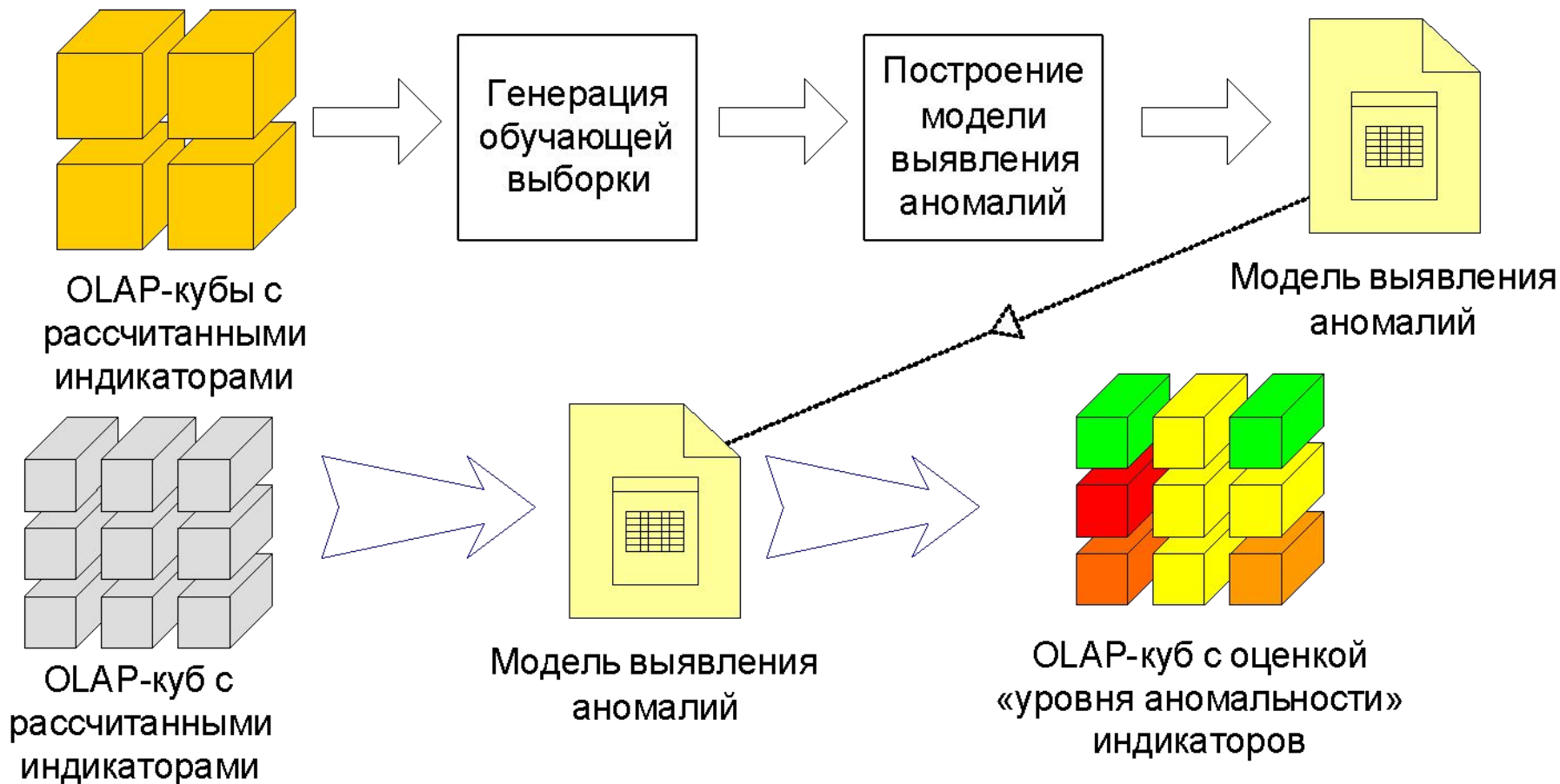


Расчет и хранение индикаторов

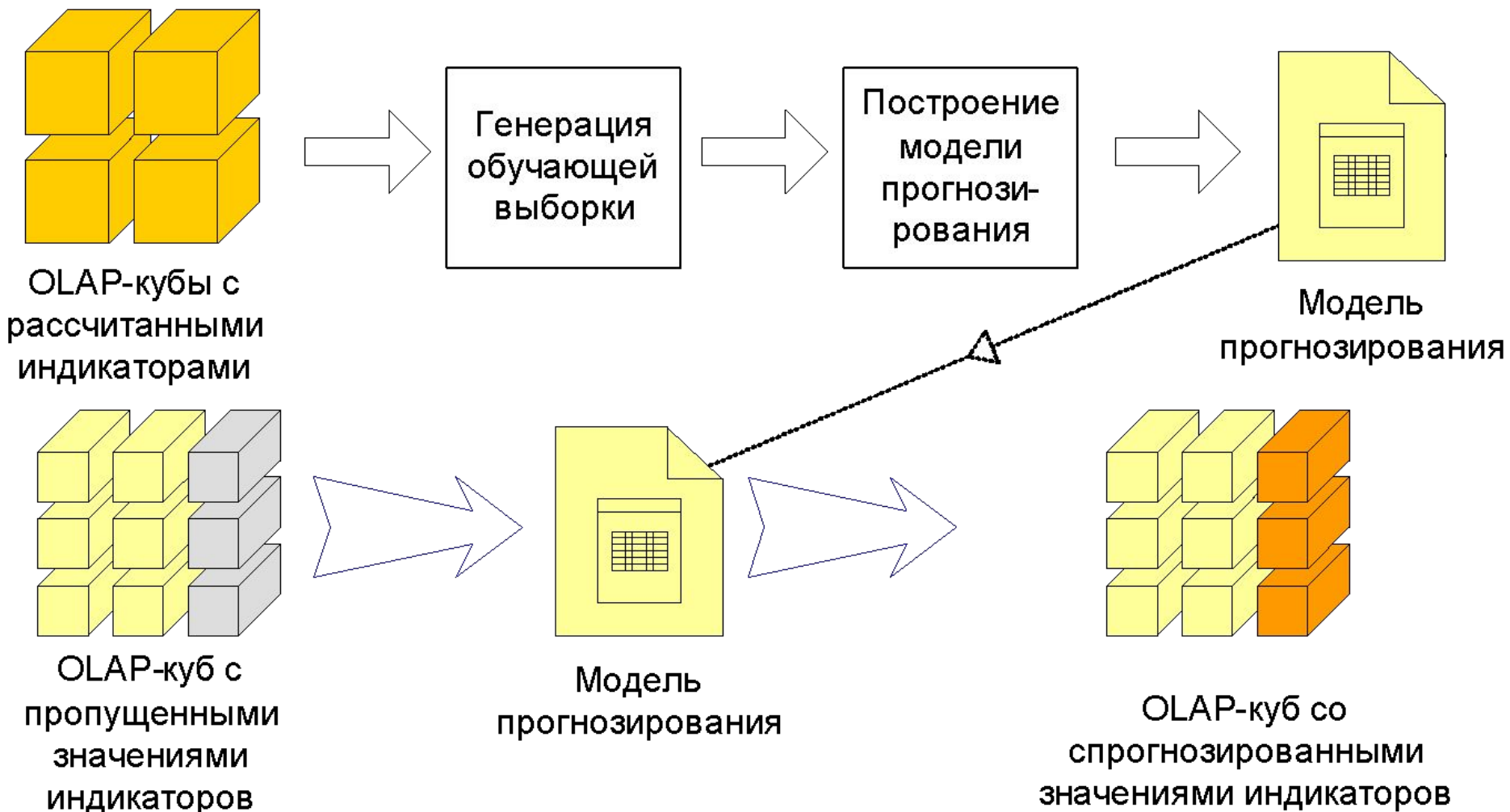
- Проведение статистического анализа и вычисление индикаторов, описывающих ситуацию



Выявление аномалий в значениях индикаторов



Определение тенденций и прогнозирование значений индикаторов



Текущие результаты

- Проектирование и создание рабочего места аналитика ситуационного центра мониторинга и анализа ситуаций:
 - Просмотр ситуации по срезам OLAP-куба в виде сводной таблицы, диаграммы или отображения на карте
 - Просмотр результатов выявления аномалий
 - Просмотр результатов прогнозирования
- Разработка и реализация специальных ИАД алгоритмов поиска аномалий и прогнозирования с учетом специфики данных – срезы OLAP куба.



Спасибо за внимание!

И

Вопросы?

д.ф.-м.н. И.В.Машечкин (mash@cs.msu.su),
к.ф.-м.н. М.И. Петровский (michael@cs.msu.su)

лаборатория «Технологий программирования»
ВМиК МГУ им. М.В. Ломоносова