

Министерство науки и образования РФ  
ГОУ ВПО «Сибирская государственная автомобильно-дорожная  
академия (СибАДИ)»

Г.А. Калачев, О.Н. Стасюк

# ИНФОРМАЦИОННО-АНАЛИТИЧЕСКИЕ СИСТЕМЫ

Учебное пособие

Омск  
СибАДИ  
2010

УДК 004.031 : 004.032.2  
ББК 32.973.233  
К 17

*Рецензенты:*

д. техн. наук, проф. Б.Н. Епифанцев (СибАДИ)  
к. техн. наук, доцент Р.А. Ахмеджанов (ОмГУПС)  
к. физ – мат. наук, М.Н. Рассказова (Институт Сервиса)

Работа одобрена редакционно - издательским советом академии в качестве учебного пособия для студентов специальностей направления «Информационная безопасность» и других категорий студентов и инженеров, занимающихся обработкой данных.

**Калачев Г.А., Стасюк О.Н.**

**К17 Информационно-аналитические системы:** учебное пособие/Г.А. Калачев, О.Н. Стасюк – Омск: Изд-во СибАДИ, 2010. – 101 с.

Данное учебное пособие разработано для изучения таких аналитических инструментов, как информационно-аналитическая система «Семантический архив» и аналитическая платформа «Deductor», которые позволяют решать задачи извлечения, очистки, анализа, моделирования и визуализации данных, прогнозирования событий и создания систем отчетности.

Пособие позволит с минимальными усилиями приобрести навыки работы с этими мощными программными средствами, а также изучить с их помощью основные методы обработки и анализа данных.

Пособие предназначено для широкого круга специалистов и прежде всего для студентов специальностей направления «Информационная безопасность».

Работа выполнена в рамках реализации федеральной целевой программы «Научные и научно-педагогические кадры инновационной России» на 2009-2013 гг, а также при поддержке ОАО «Аналитические бизнес решения» и кафедры информационная безопасность СибАДИ.

Ил. 87. Библиогр.: 11 назв.

© ГОУ «СибАДИ», 2010

*Учебное издание*

Григорий Александрович Калачев,  
Ольга Николаевна Стасюк

ИНФОРМАЦИОННО-АНАЛИТИЧЕСКИЕ  
СИСТЕМЫ

Учебное пособие

\*\*\*

Редактор Н.И. Косенкова

\*\*\*

Подписано к печати .09 . 2010  
Формат 60×90 1/16. Бумага писчая  
Оперативный способ печати  
Гарнитура Times New Roman  
Усл. п. л. 6,25 , уч.-изд. л. 4,59  
Тираж 100 экз. Заказ № \_\_\_\_  
Цена договорная

Издательство СибАДИ  
644099, г. Омск, ул. П. Некрасова, 10  
Отпечатано в подразделении ОП издательства СибАДИ

## СОДЕРЖАНИЕ

Введение.....	4
1. Аналитическая платформа «Deductor».....	7
1.1. Описание платформы.....	7
1.2. Возможности платформы.....	7
1.3. Состав системы.....	9
2. Комплекс лабораторных работ по изучению возможностей аналитической платформы «Deductor».....	12
Лабораторная работа №1. Знакомство с АП «Deductor».....	12
Лабораторная работа №2. Реализация алгоритма построения деревьев решений.....	16
Лабораторная работа №3. Логистическая регрессия и ROC- анализ.....	24
Лабораторная работа №4. Применение алгоритма кластеризации: самоорганизующиеся карты Кохонена.....	32
Лабораторная работа №5. Поиск ассоциативных правил.....	40
3. Информационно-аналитическая система «Семантический архив».....	46
3.1. Общее описание системы.....	46
3.2. Возможности системы.....	46
3.3. Технология работы системы.....	47
3.4. Типы автоматизированных рабочих мест (АРМ).....	47
3.5. Технологический цикл работы.....	48
3.6. Технология обработки документов.....	48
3.7. Технология описания факта.....	49
3.8. Основные функции системы.....	49
3.9. Отличительные особенности системы.....	50
3.10. Цели внедрения системы.....	50
3.11. Пользователи системы.....	51
4. Комплекс лабораторных работ по изучению информационно- аналитической системы «Семантический архив».....	52
Лабораторная работа №1. Сценарий работы пользователя с модулем поиска «Искатель».....	52
Лабораторная работа №2. Добавление данных в базы данных.....	61
Лабораторная работа №3. Работа в витрине «сквозного поиска».....	67
Лабораторная работа №4. Перенос данных из АРМ «Оператор» в «Аналитик».....	77
Лабораторная работа №5. Построение семантических сетей.....	84
Лабораторная работа №6. Формирование дайджестов статей.....	90
Библиографический список.....	99

## ВВЕДЕНИЕ

Информационно-аналитические системы получают все более широкое распространение в бизнесе, поскольку позволяют принимать качественные решения.

Действительно, чтобы не отстать от ритма жизни, который задает современный бизнес, руководителям, топ-менеджерам, принимающим решения, требуется полная и всеобъемлющая информация о текущем положении дел, состоянии рынка, позициях конкурентов. Причем крайне важно получать такую информацию своевременно. Однако на практике, решая управленческие задачи, руководство сталкивается с рядом серьезных трудностей, прежде всего связанных со сбором, подготовкой и дальнейшим использованием данных.

Сбор, обработка и последующий анализ данных требует значительных временных и интеллектуальных затрат. Иметь в штате несколько аналитиков - очень дорогостоящее мероприятие, а использование аутсорсинга не может гарантировать абсолютную безопасность бизнеса, поскольку придется раскрывать коммерческую информацию.

Чтобы обеспечить полноценный анализ состояния организации и получить достоверные прогнозы, нужно собрать, сохранить, обработать и проанализировать огромное количество сведений из самых разнообразных источников, а для этого необходим инструмент, объединяющий в себе передовые информационные технологии и развитый математический аппарат. Поэтому роль специализированных информационно-аналитических систем (ИАС) в последнее время становится все более значительной. В ряде случаев они даже способны предложить руководителю качественные трактовки результатов, полученных в ходе применения аналитических методик, формируя автоматизированные аналитические записки.

До 2000<sup>1</sup> года господствующее положение на этом рынке занимали программные продукты иностранных фирм. В настоящее время положение дел меняется – появилось несколько российских программ такого назначения, по своим характеристикам способных успешно конкурировать с зарубежными, а по ряду параметров и превосходящие их. Главные преимущества российских продуктов

---

<sup>1</sup> По данным информационного бизнес - портала [www.market-pages.ru](http://www.market-pages.ru)

видны в соотношении цена/качество, отсутствие проблем локализации и пр.

К подобным информационно-аналитическим инструментам можно отнести ИАС «Семантический архив» и аналитическую платформу (АП) «Deductor», которые предназначены для решения большого спектра задач и способны выполнять следующие функции:

- обеспечение своевременного поступления надежной и всесторонней информации по интересующим вопросам;
- описание сценария действий конкурентов, которые могут затрагивать текущие интересы организации;
- осуществление постоянного мониторинга событий во внешней конкурентной среде и на рынке, которые могут иметь значение для интересов организации и системы защиты информации;
- обеспечение безопасности собственных информационных ресурсов;
- обеспечение эффективности сбора, анализа и распространения информации, исключение дублирования исходных и производственных данных;
- обеспечение эффективной обработки поступающей информации и возможности моделирования событий;
- возможность прогнозирования развития событий;
- управление рисками и др.

Эффективное применение этих средств - один из факторов выживаемости и успеха предприятия в условиях острой конкурентной борьбы. Однако необходимо отметить следующее обстоятельство: такие программные продукты весьма дороги и пока малодоступны для массового потребителя. А это не позволяет повсеместно использовать их для обучения студентов, а отсутствие практики плохо сказывается на усвоении материала. Кроме того, на рынке очень мало пособий по анализу данных с использованием АП «Deductor», а учебников по ИАС «Семантический архив» нет вообще.

Учитывая эти факты, было разработано учебное пособие для изучения двух описанных выше отечественных аналитических систем.

Пособие позволяет с минимальными усилиями приобрести навыки работы с этими мощными программными средствами, а также изучить с их помощью основные методы обработки и анализа данных.

Учебное пособие включает в себя как теоретические разделы, так и практические рекомендации по решению аналитических задач на

примере выполнения комплекса лабораторных работ. Лабораторные работы включают в себя: цель, теоретическое описание, практические рекомендации по выполнению работы, задания для закрепления материала. В учебнике наглядно показана работа с ИАС: интерфейс, последовательные диалоговые окна.

Пособие, прежде всего, предназначено для студентов специальностей направления «Информационная безопасность», желающим получить основы работы со средствами аналитической разведки.

Разделы учебника будут полезны преподавателям при проведении лабораторных работ, поскольку содержат достаточно теории для выполнения лабораторных работ и не требуют поиска дополнительной литературы и подготовки. Работа рассчитана также на широкий круг специалистов и инженеров, занимающихся обработкой данных.

# **1. АНАЛИТИЧЕСКАЯ ПЛАТФОРМА «DEDUCTOR»**

## **1.1. Описание платформы**

Аналитическая платформа (АП) – это комплекс программных продуктов, связанных единой архитектурой. АП относятся к группе программных продуктов и технологий под общим названием Business Intelligence<sup>2</sup> и автоматизируют функции анализа бизнеса и поддержки принятия решений. Подобные системы стали появляться на мировом рынке информационных технологий в 80-90 годах прошлого столетия.

АП «Deductor», разработанная компанией BaseGroup Labs, является одной из лучших отечественных разработок в данной области.

Технологии и методики анализа данных, реализованные в этой платформе, позволяют на базе единой архитектуры пройти все этапы построения аналитической системы: начиная от создания хранилища данных и заканчивая построением моделей.

В ней сосредоточены самые современные методы извлечения, очистки, манипулирования и визуализации данных. С применением АП «Deductor» становятся доступными механизмы моделирования, прогнозирования, кластеризации, поиска закономерностей и многие другие технологии обнаружения знаний (Knowledge Discovery in Databases), добычи данных (Data Mining) и многомерного анализа (OLAP).

## **1.2. Возможности платформы**

«Deductor» предназначен для решения широкого спектра задач, прикладная область значения не имеет, т.к. механизмы, реализованные в АП, с успехом применяются на финансовых рынках, в страховании, торговле, телекоммуникациях, промышленности, медицине, маркетинге и других сферах деятельности.

Рассмотрим наиболее популярные задачи, решаемые при помощи «Deductor»:

---

<sup>2</sup> Это процесс извлечения многоаспектной информации и превращение её в знания для эффективного управления бизнесом, осуществляемый конечными пользователями с помощью специальных технологий, методов и средств.



– *Создание систем отчетности.* Содержащиеся в хранилище данные можно просматривать, используя различные визуализаторы, например, OLAP кубы, таблицы, диаграммы, гистограммы.

– *Data Mining проекты.* Data Mining переводится как «добыча» или «раскопка данных». Это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных областях человеческой деятельности. Он может применяться везде, где возникает потребность в глубоком анализе данных, но чаще всего речь идет об анализе коммерческой информации.

Некоторые задачи, решаемые при помощи методов Data Mining:

- анализ и управление рисками;
- сегментация клиентов, продуктов, услуг;
- определение особенностей поведения клиентов;
- промышленная диагностика, обнаружение источников и причин возникновения дефектов;
- идентификация критических ситуаций;
- оценка кредитоспособности физических и юридических лиц и многое другое.

– *Механизмы очистки данных.* На практике исходные данные чаще всего бывают «сырыми». Очищенные данные содержат наиболее ценную для анализа информацию, из которой исключены противоречивые и дублирующиеся данные, устранены аномалии и шумы. Во многих случаях достаточно провести только очистку данных, и выводы будут очевидны. Кроме того, очистка данных позволяет получить лучшие результаты при дальнейшем построении моделей.

– *Прогнозирование.* Это одна из самых востребованных задач анализа. В Deductor включены механизмы построения прогностических моделей, в том числе с использованием самообучающихся алгоритмов. Достаточно построить модель, прогнозирующую изменение на 1 шаг, и автоматически использовать ее на произвольное количество отсчетов вперед. Это позволяет получать качественные прогнозы, способные подстраиваться под изменяемую ситуацию.

– *Моделирование.* Построение моделей – универсальный способ анализа. В большинстве случаев при исследовании процесса или объекта мы строим его модель, но не всегда эта модель

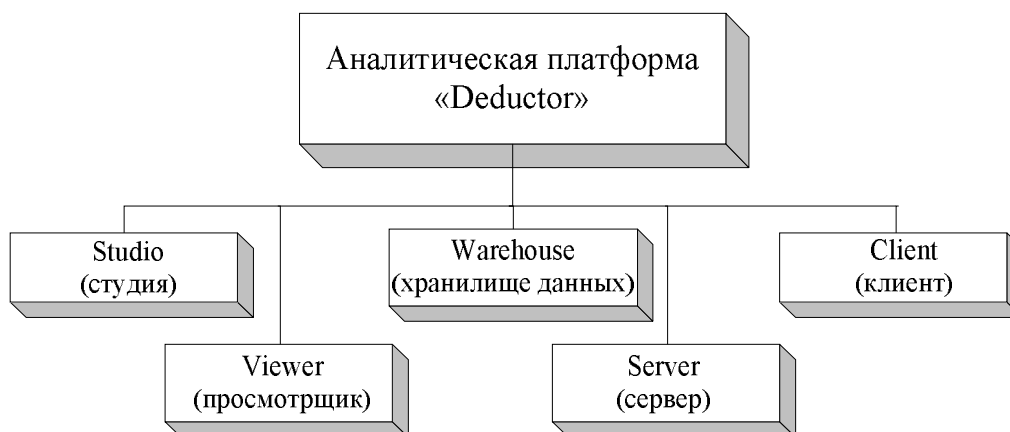
формализована, т.е. описана таким образом, чтобы ею мог воспользоваться другой человек, подавая свои входные данные и получая результат.

В Deductor основной акцент сделан на самообучающиеся методы и машинное обучение. Такие алгоритмы являются универсальными, решающими большой спектр задач, и при этом просты в применении. Полученные результаты можно просмотреть в виде таблиц, кубов, карт, деревьев и прочее.

– *Анализ «Что, если...?»*. При принятии управленческих решений полезным инструментом является сценарное моделирование «Что, если...?», которое позволяет моделировать будущие показатели деятельности с учетом имеющихся взаимосвязей. В частности, моделирование «Что, если...?» предназначено для анализа влияния исходных показателей на целевой показатель. Для реализации этого механизма в Deductor существует специальный визуализатор. При этом способ построения модели значения не имеет, работа со всеми алгоритмами производится одинаково. Результаты анализа можно просмотреть как в табличном, так и графическом виде.

### 1.3. Состав системы

«Deductor» состоит из пяти частей, показанных ниже на рисунке.



Состав АП «Deductor»

*Deductor Studio* является аналитическим ядром всей платформы, основанном на работе следующих механизмов:

- мастера импорта исходного набора данных;
- мастера обработки;

- мастера визуализации;
- мастера экспорта полученных данных.

То есть реализованные в Studio механизмы позволяют в рамках единого процесса пройти весь цикл анализа данных – получить исходную информацию из стороннего источника (хранилище данных Deductor Warehouse, промышленные СУБД, текстовые файлы, офисные приложения, 1С:Предприятие и прочее), обработать эту информацию, отобразить полученные результаты разными способами (например, с помощью таблиц, диаграмм, карт и т.п.) и экспортировать их (форматы Microsoft Excel, Microsoft Word, HTML, XML и др.).

Используется аналитиками-экспертами, которые занимаются созданием сценариев обработки.

*Deductor Warehouse* - это хранилище данных, содержащее в себе всю информацию, необходимую для анализа конкретной предметной области.

Использование Deductor Warehouse позволяет обеспечить:

- удобный доступ к данным;
- высокую скорость их обработки;
- непротиворечивость информации;
- централизованное хранение;
- поддержку процесса анализа данных.

От пользователя не требуется знание структуры хранения данных и языка запросов, для начала работы с хранилищем необходимо вызвать мастер импорта и выбрать интересующую его информацию.

Загрузка данных в Warehouse производится при помощи Deductor Studio либо Server. Хранилище строится на базе одной из трех СУБД: Oracle, MS SQL или Firebird.

*Deductor Client* – это динамическая библиотека для удаленной работы с Deductor Server. Обеспечивает доступ к серверу из сторонних приложений и управление его работой.

*Deductor Viewer* является рабочим местом конечного пользователя. У пользователей должна быть возможность просмотра полученных данных и настройки способа отображения, возможность экспорта их в офисные программы и другие форматы, а также им должна быть доступна функция печати необходимых сведений.

Еще одна важная функция Viewer – разграничения прав доступа к данным, т.е. пользователь может получить информацию, разрешенную ему в виде отчетов, при этом, не имея прав на получение новых данных.

*Deductor Server* – это служба, позволяющая осуществлять удаленную аналитическую обработку данных.

Использование сервера является оптимальным для корпоративной среды, что подтверждено следующими преимуществами:

- простота интеграции, т.к. удаленный доступ к нему осуществляется с помощью специальной бесплатно распространяемой библиотеки *Deductor Client*;

- удаленный доступ позволяет взаимодействовать с сервером как в локальной сети, так и через Интернет (по протоколу TCP/IP). Также *Server* позволяет автоматически восстанавливать соединения с базами данных при их обрыве и отключать «зависшие» сессии;

- высокая производительность и оптимальное использование оборудования достигается за счет многопоточной обработки, встроенных механизмов балансировки нагрузки, переноса на высокопроизводительный сервер наиболее ресурсоемких операций, кэширования загруженных ранее проектов, что также значительно повышает скорость аналитической обработки;

- удобство администрирования достигается за счет функции протоколирования, позволяющей фиксировать ход выполнения работ, возникающие ошибки, анализировать причины возникновения ошибок и производительность системы. Кроме того, в *Deductor Server* встроен планировщик заданий.

«*Deductor*» является идеальной платформой для создания систем поддержки принятий решений.

В «*Deductor*» используются мощные технологии анализа данных, но при этом акцент сделан на самообучающиеся методы, что позволяет строить системы, способные реагировать на изменение ситуации.

Объединение всех описанных выше механизмов в рамках единой программы обеспечивает принципиально новое качество – уменьшается время создания законченных решений, упрощается интеграция с другими приложениями, увеличивается производительность. Все это сочетается с гибкостью и простотой использования.

Платформа позволяет минимизировать требования к обучению персонала, поскольку все необходимые операции выполняются автоматически при помощи подготовленных ранее сценариев обработки.

За основу представленного далее комплекса лабораторных работ взят практикум, разработанный А.А. Барсегяном<sup>3</sup>. Практикум был переработан и дополнен (проработаны разделы: теоретические основы и задания для повторения). Все лабораторные работы выполнены с использованием «Deductor Academic»<sup>4</sup> версии 5.1.

## **2. КОМПЛЕКС ЛАБОРАТОРНЫХ РАБОТ ПО ИЗУЧЕНИЮ ВОЗМОЖНОСТЕЙ АНАЛИТИЧЕСКОЙ ПЛАТФОРМЫ «DEDUCTOR»**

### **Лабораторная работа №1. Знакомство с АП «Deductor»**

#### *1.1. Основная цель*

Целью выполнения данной лабораторной работы является:

- получение первоначальных сведений о возможностях аналитической платформы;
- изучение основных модулей; работа с мастерами импорта, экспорта, обработки и визуализации данных.

#### *1.2. Теоретическая часть*

АП «Deductor» применима для решения большого спектра задач, таких как создание аналитической отчетности, прогнозирование, поиск закономерностей и пр. Можно сказать, что данная система применима в задачах, где требуется консолидация и отображение данных различными способами, построение моделей и последующее применение полученных моделей к новым данным.

Рассмотрим некоторые задачи, решаемые АП:

– *Системы корпоративной отчетности.* Готовое хранилище данных и гибкие механизмы предобработки, очистки, загрузки, визуализации позволяют быстро создавать законченные системы отчетности в сжатые сроки.

– *Обработка нерегламентированных запросов.* Конечный пользователь может с легкостью получить ответ на вопросы типа "Сколько было продаж товара по группам в Московскую область за

---

<sup>3</sup> Методы и модели анализа данных: OLAP и Data Mining /А.А. Барсегян, М.С.

Куприянов, В.В. Степаненко, И.И. Холод – СПб.: БХВ-Петербург, 2004.- 336 с.: ил.

<sup>4</sup> Бесплатная версия предназначена для образовательных целей.

прошлый год с разбивкой по месяцам?" и просмотреть результаты наиболее удобным для него способом.

– *Анализ тенденций и закономерностей, планирование, ранжирование.* Простота использования и интуитивно понятная модель данных позволяет вам проводить анализ по принципу «Что, если...?», соотносить ваши гипотезы со сведениями, хранящимися в базе данных, находить аномальные значения, оценивать последствия принятия бизнес-решений.

– *Прогнозирование.* Построив модель на исторических примерах, вы можете использовать ее для прогнозирования ситуации в будущем. По мере изменения ситуации нет необходимости перестраивать все, необходимо всего лишь дообучить модель.

– *Управление рисками.* Реализованные в системе алгоритмы дают возможность достаточно точно определиться с тем, какие характеристики объектов и как влияют на риски, благодаря чему можно прогнозировать наступление рискованного события и заблаговременно принимать необходимые меры к снижению размера возможных неблагоприятных последствий.

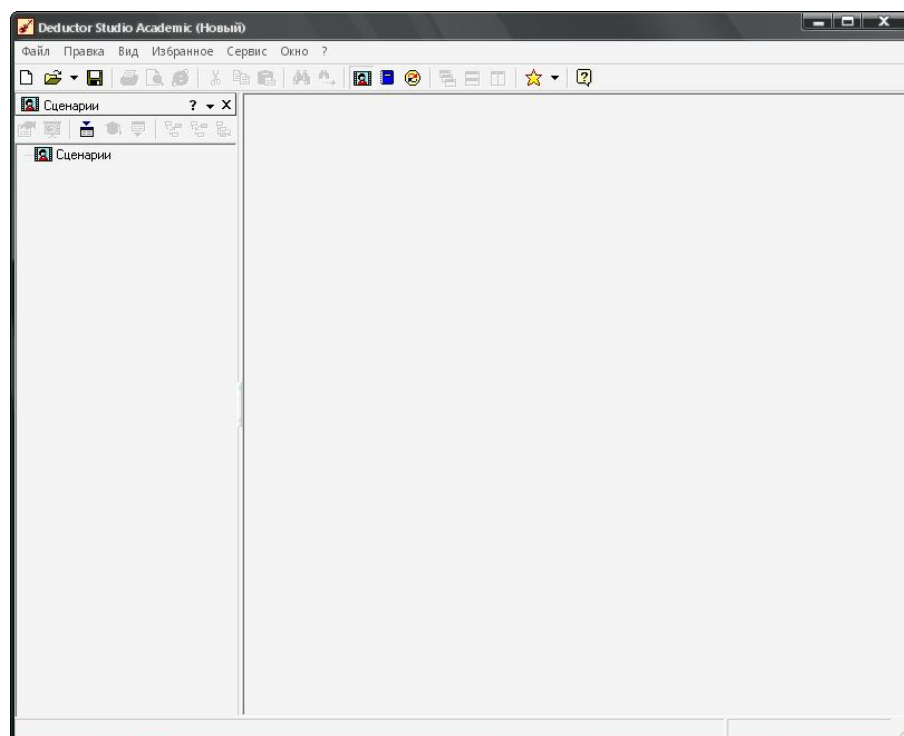
– *Анализ данных маркетинговых и социологических исследований.* Анализируя сведения о потребителях, можно определить, кто является вашим клиентом и почему. Как изменяются их пристрастия в зависимости от возраста, образования, социального положения, материального состояния и множества других показателей. Понимание этого будет способствовать правильному позиционированию ваших продуктов и стимулированию продаж.

– *Диагностика.* Механизмы анализа, имеющиеся в системе Deductor, с успехом применяются в медицинской диагностике и диагностике сложного оборудования. Например, можно построить модель на основе сведений об отказах. При ее помощи быстро локализовать проблемы и находить причины сбоев.


– *Обнаружение объектов на основе нечетких критериев.* Часто встречается ситуация, когда необходимо обнаружить объект, основываясь не на таких четких критериях, как стоимость, технические характеристики продукта, а на размытых формулировках, например, найти продукты, похожие на ваши с точки зрения потребителя.

### 1.3. Практическая часть

После запуска «Deductor Studio Academic» появится главное окно программы.



Главное окно после запуска программы  
Deductor Studio

Для начала работы необходимо создать новый сценарий, воспользуемся для этого *мастером импорта* (кнопка  в левой части главного окна либо клавиша F6).

Импорт данных включает в себя:


- выбор типа источника данных;
- выбор файла источника данных;
- указание параметров импорта;
- указание параметров столбцов;
- выбор способа отображения данных (при выборе «Диаграммы», «Гистограммы» или «OLAP-куба» потребуется дополнительно указать параметры построения);
- указание имени, метки и описания данных.

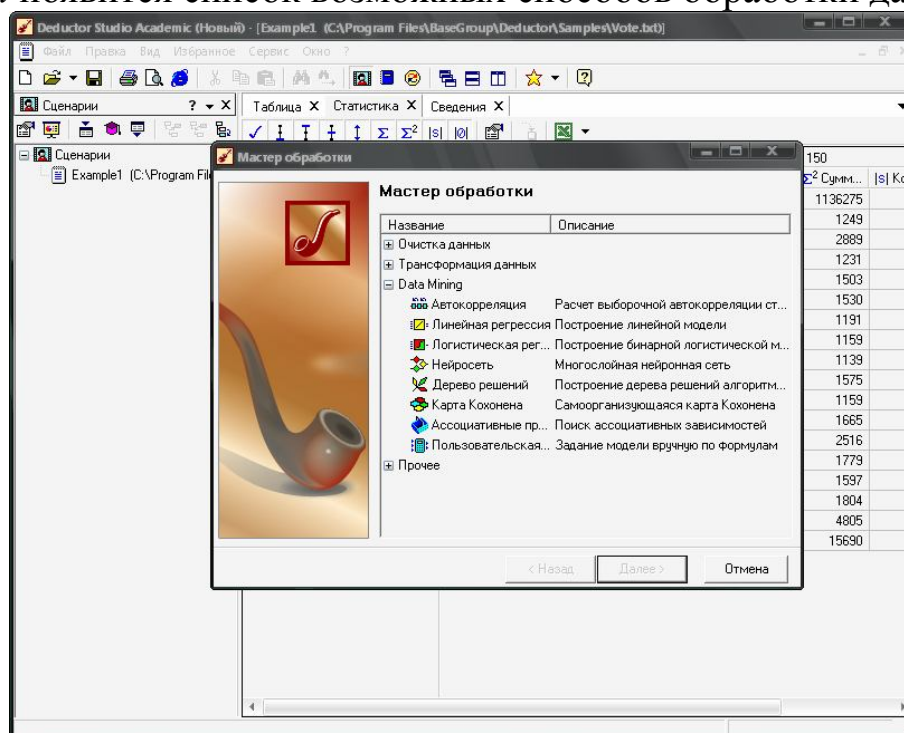
Выполнив вышеуказанные действия по импорту данных, на панели «Сценарии» мы получим новый узел, с заданными именем, меткой и описанием.

Статистика: Кол-во значений = 150

Метка столбца	Мини...	Макс...	Сред...	Стан...	Сумма	Σ <sup>2</sup> Сумм...	s  Кол
1 9.0 Код	1	150	75,5	3679924569	11325	1136275	
2 ab Проект по инвалидам	2	11	2,673	1,09	401	1249	
3 ab Проект по водным ре...	2	11	3,42	2,759	513	2889	
4 ab Проект по усыновлен...	2	11	2,553	1,303	383	1231	
5 ab Закон о врачах	2	11	2,82	1,443	423	1503	
6 ab Проект по Сальвадору	2	11	2,76	1,612	414	1530	
7 ab Закон о религиях	2	11	2,5	1,304	375	1191	
8 ab Антиспутниковый про...	2	11	2,553	1,102	383	1159	
9 ab Проект помощи Ника...	2	11	2,527	1,103	379	1139	
10 ab Проект по ракетам	2	11	2,82	1,601	423	1575	
11 ab Закон об иммигрантах	2	11	2,553	1,102	383	1159	
12 ab Проект по альтернат...	2	11	2,94	1,573	441	1665	
13 ab Закон об образовании	2	11	3,307	2,425	496	2516	
14 ab Проект по фондам	2	11	2,9	1,864	435	1779	
15 ab Проект по преступно...	2	11	2,753	1,757	413	1597	
16 ab Проект по таможенн...	2	11	2,933	1,856	440	1804	
17 ab Проект по экспорту	2	11	4,247	3,754	637	4805	
18 ab Класс	8	13	9,933	2,443	1490	15690	

Пример создания сценария, вкладка «Статистика»


Изучим возможности *мастера обработки* (кнопка  в левой части главного окна либо клавиша F7). После запуска *мастера обработки* появится список возможных способов обработки данных.




Список доступных способов обработки данных



Все способы разделены на четыре основные группы: очистка данных, трансформация данных, Data Mining, пр. Каждый способ обработки имеет название и краткое описание. Выбор способа зависит от целей обработки данных (например, сортировка и фильтрация данных, построение дерева решений и пр.).

*Мастер визуализации* позволяет определить способ отображения данных, указать метки и добавить описание к проекту. Запустить его можно с помощью кнопки  либо клавишей F5.

Готовый проект можно экспортировать, воспользовавшись *мастером экспорта* (кнопка  основного окна либо клавиша F8). Указав параметры, проект можно перенести в один из доступных форматов.

### *1.4. Задание*

1. Опишите назначение и возможности АП «Deductor».
2. Запустите программу «Deductor Studio Academic», ознакомьтесь с назначением кнопок и контекстным меню главного окна программы.
3. Воспользуйтесь *мастером импорта* данных (импортируйте любой файл, например из C:\Program Files\ BaseGroup\ Deductor\ Samples\ \*.txt ).
4. Ознакомьтесь с доступными способами обработки данных.
5. Изучите возможности *мастера визуализации* и *экспорта*. Какие параметры доступны для *мастера экспорта* данных?
6. Создайте отчет.

## **Лабораторная работа №2. Реализация алгоритма построения дерева решений**

### *2.1. Основная цель*

Изучить алгоритм «Построение дерева решений» и научиться обрабатывать с его помощью данные.

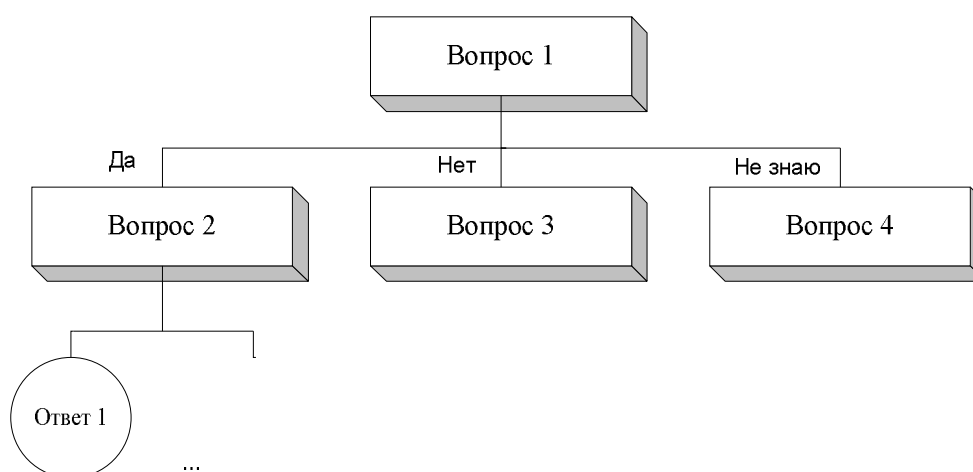
### *2.2. Теоретическая часть*

Своевременная разработка и принятие правильного решения - это одна из главных задач работы управленческого персонала

организации, т.к. необдуманное решение может дорого обойтись компании. Зачастую на практике результат одного решения заставляет нас принимать следующее решение и т. д. Когда же нужно принять несколько решений в условиях неопределенности, когда каждое решение зависит от исхода предыдущего, то применяют схему, называемую деревом решений.

Дерево решений это графическое изображение процесса принятия решений, в котором отражены альтернативные решения, соответствующие вероятности, и выигрыши для любых комбинаций альтернатив.

Дерево решений представляет один из способов разбиения множества данных на классы или категории. Корень дерева неявно содержит все классифицируемые данные, а листья определенные классы после выполнения классификации. Промежуточные узлы дерева представляют пункты принятия решения о выборе.



Структура дерева решений

### *Построение дерева решений*

Пусть нам задано некоторое обучающее множество  $T$ , содержащее объекты, каждый из которых характеризуется  $m$  атрибутами, причем один из них указывает на принадлежность объекта к определенному классу.

Пусть через  $\{C_1, C_2, \dots, C_k\}$  обозначены классы, тогда существуют 3 ситуации:

- множество  $T$  содержит один или более примеров, относящихся к одному классу  $C_k$ . Тогда дерево решений для  $T$  – это лист, определяющий класс  $C_k$ ;

– множество  $T$  не содержит ни одного примера, т.е. пустое множество. Тогда это снова лист, и класс, ассоциированный с листом, выбирается из другого множества отличного от  $T$ , скажем, из множества, ассоциированного с родителем;

– множество  $T$  содержит примеры, относящиеся к разным классам. В этом случае следует разбить множество  $T$  на некоторые подмножества. Для этого выбирается один из признаков, имеющий два и более отличных друг от друга значений  $O_1, O_2, \dots O_n$ .  $T$  разбивается на подмножества  $T_1, T_2, \dots T_n$ , где каждое подмножество  $T_i$  содержит все примеры, имеющие значение  $O_i$  для выбранного признака. Эта процедура будет рекурсивно продолжаться до тех пор, пока конечное множество не будет состоять из примеров, относящихся к одному и тому же классу.

Вышеописанная процедура лежит в основе многих современных алгоритмов построения дерева решений, этот метод известен еще под названием «разделение и захват». Очевидно, что при использовании данной методики построение дерева решений будет происходить сверху вниз.

### *Области применения дерева решений*

Дерево решений является прекрасным инструментом в системах поддержки принятия решений, интеллектуального анализа данных (Data Mining). В областях, где высока цена ошибки, они послужат отличным подспорьем аналитика или руководителя.

Дерево решений успешно применяется для решения практических задач в следующих областях:

– *Банковское дело.* Оценка кредитоспособности клиентов банка при выдаче кредитов.

– *Промышленность.* Контроль качества продукции (выявление дефектов), испытания без разрушений (например, проверка качества сварки) и т.д.

– *Медицина.* Диагностика различных заболеваний.

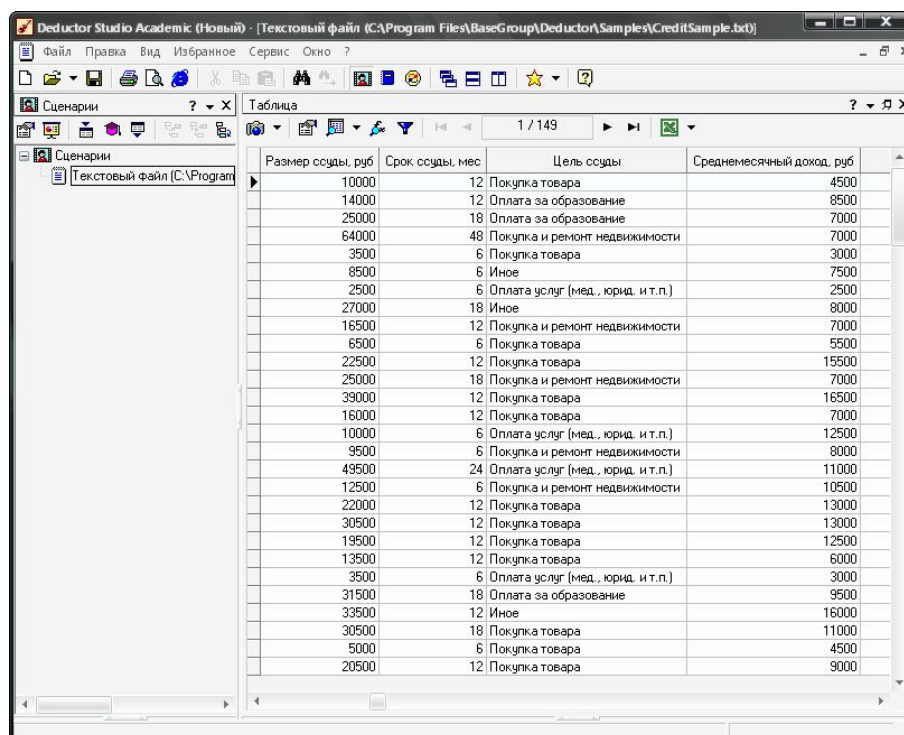
– *Молекулярная биология.* Анализ строения аминокислот.

Это далеко не полный список областей, где можно использовать дерево решений, т.к. еще многие потенциальные области применения не исследованы.

### 2.3. Практическая часть

Для загрузки данных примера импортируйте файл C:\Program Files\BaseGroup\Deductor\Samples\CreditSample.txt в АП «Deductor» с помощью *мастера импорта*. Все параметры импорта примите установленными по умолчанию. В окне выбора способа отображения данных выберите «Таблица», если он не выбран по умолчанию.

В результате в основном окне появится таблица, заполненная из указанного файла.

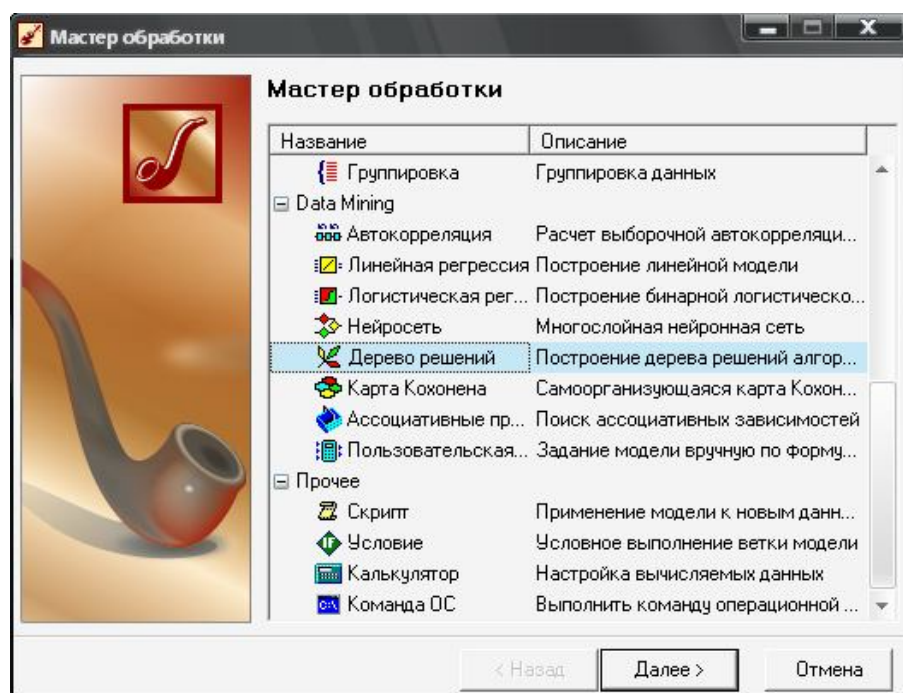


The screenshot shows the 'Deductor Studio Academic' interface. The main window displays a table with 4 columns: 'Размер ссуды, руб' (Loan amount, rub), 'Срок ссуды, мес' (Loan term, months), 'Цель ссуды' (Loan purpose), and 'Среднемесячный доход, руб' (Average monthly income, rub). The table contains 20 rows of data. The left sidebar shows a tree view with 'Сценарии' (Scenarios) and 'Текстовый файл [C:\Program Files\BaseGroup\Deductor\Samples\CreditSample.txt]' selected. The top menu bar includes 'Файл', 'Правка', 'Вид', 'Избранное', 'Сервис', and 'Окно'.

Размер ссуды, руб	Срок ссуды, мес	Цель ссуды	Среднемесячный доход, руб
10000	12	Покупка товара	4500
14000	12	Оплата за образование	8500
25000	18	Оплата за образование	7000
64000	48	Покупка и ремонт недвижимости	7000
3500	6	Покупка товара	3000
8500	6	Иное	7500
2500	6	Оплата услуг (мед., юрид. и т.п.)	2500
27000	18	Иное	8000
16500	12	Покупка и ремонт недвижимости	7000
6500	6	Покупка товара	5500
22500	12	Покупка товара	15500
25000	18	Покупка и ремонт недвижимости	7000
39000	12	Покупка товара	16500
16000	12	Покупка товара	7000
10000	6	Оплата услуг (мед., юрид. и т.п.)	12500
9500	6	Покупка и ремонт недвижимости	8000
49500	24	Оплата услуг (мед., юрид. и т.п.)	11000
12500	6	Покупка и ремонт недвижимости	10500
22000	12	Покупка товара	13000
30500	12	Покупка товара	13000
19500	12	Покупка товара	12500
13500	12	Покупка товара	6000
3500	6	Оплата услуг (мед., юрид. и т.п.)	3000
31500	18	Оплата за образование	9500
33500	12	Иное	16000
30500	18	Покупка товара	11000
5000	6	Покупка товара	4500
20500	12	Покупка товара	9000

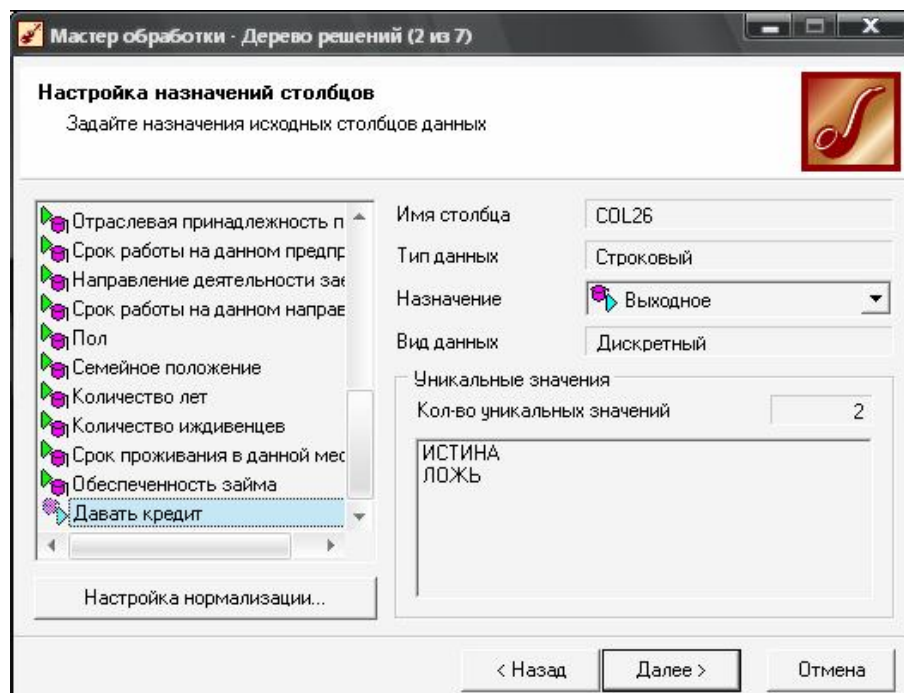
Итог импорта данных

Запустите *мастер обработки данных*. В появившемся окне в разделе Data Mining выберите метод обработки «Дерево решений» и нажмите «Далее».



Мастер обработки данных

На вкладке «Настройка значения столбцов» необходимо задать назначения столбцов данных. Почти все столбцы автоматически получили значение «Входные». Значение поля «Выдать кредит», которое принимает только два значения «Да» или «Нет», необходимо установить в «Выходное». Также необходимо обозначить столбцы «Код» и «№ паспорта» как «Неиспользуемые» (так как значения этих столбцов уникальны, а это не позволит их классифицировать).

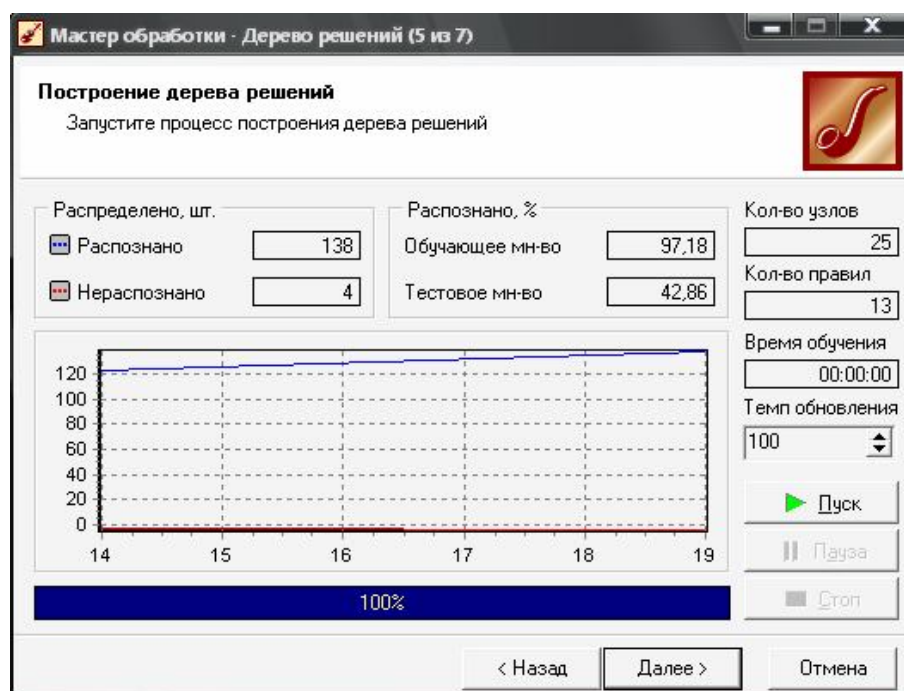


Окно настройки назначений столбцов

Далее следует окно настройки разбиения исходного множества данных на подмножества. Оставьте это окно без изменений и нажмите кнопку «Далее».

Следующий этап – настройка параметров обучения дерева решений. Необходимо учитывать, что чем больше значение параметра «Уровень доверия, используемый при отсечении узлов дерева», тем больше будет дерево решений в итоге.

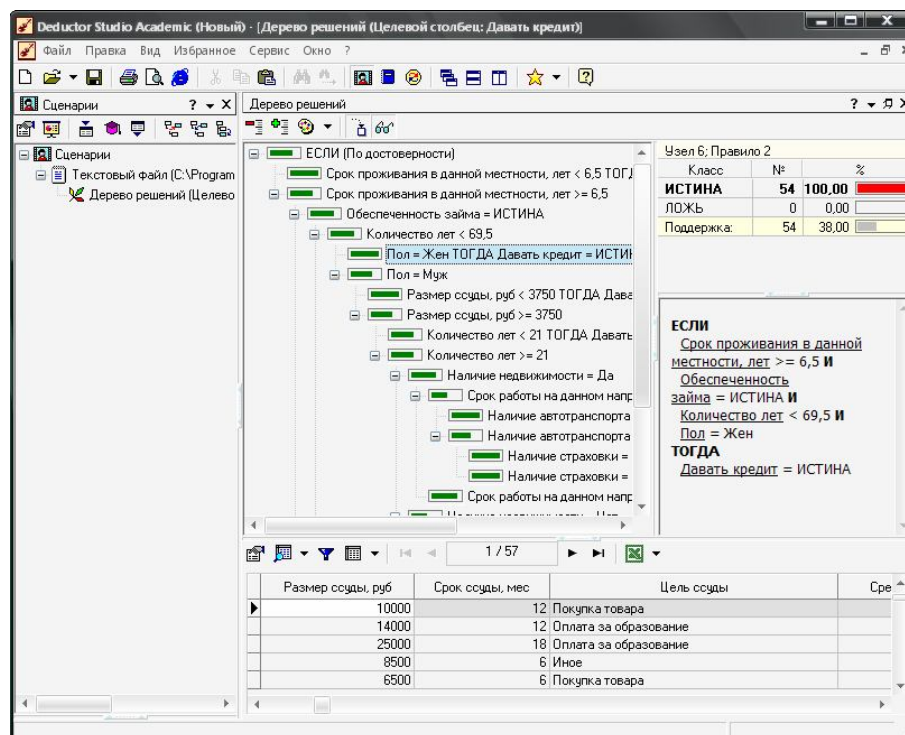
С помощью кнопки «Пуск» запускаем процесс построения дерева решений. По окончании процесса вы увидите график, отображающий уровень распознавания данных, количество узлов созданного дерева и правил, полученных в результате обработки.



Процесс построения дерева решений

В последующем окне выбора способа отображения данных выберите «Дерево решений». А в последнем окне *мастера обработки*, по желанию, укажите имя и метку.

Результатом всех вышеописанных действий будет построенное дерево решений, которое отобразится в основном окне программы. На основании этого метода можно ответить на вопрос «Давать ли человеку кредит и если да, то при каких условиях».



Готовое дерево решений

Из полученного дерева можно вывести правила выдачи кредитов. Например:

- Если срок проживания в данной местности меньше 6,5 лет, то кредит не давать.
- Если срок проживания в данной местности больше 6,5 лет, займ обеспечен, возраст больше 20,5 лет, не имеется недвижимость, но имеется банковский счет, то кредит давать.

## 2.4. Задание

1. Постройте дерево решения для описанного выше примера. Попробуйте использовать различные значения параметров обучения дерева решения и сравните полученные деревья.
2. Выведите 5 правил из построенного дерева решений.
3. Приведите 4-5 примеров, для которых можно использовать метод обработки дерева решений, реализуйте один из них.
4. Составьте отчет.



## Лабораторная работа №3. Логистическая регрессия и ROC-анализ

### 3.1. Основная цель

Научиться обрабатывать данные и прогнозировать события, используя возможности логистической регрессии и ROC-анализ.

### 3.2. Теоретическая часть

*Логистическая регрессия* — метод построения линейного классификатора, позволяющий оценивать апостериорные вероятности принадлежности объектов классам.

Вообще, регрессионная модель предназначена для решения задач предсказания значения непрерывной зависимой переменной, при условии, что эта зависимая переменная может принимать значения на интервале от 0 до 1. В силу такой специфики ее часто используют для предсказания вероятности наступления некоторого события в зависимости от значений некоторого числа предикторов.

При изучении линейной регрессии мы исследуем модели вида

$$y = a + b_1x_1 + b_2x_2 + \dots + b_nx_n.$$

Здесь зависимая переменная  $y$  является непрерывной, и мы определяем набор независимых переменных  $x_i$  и коэффициенты при них  $b_i$ , которые позволили бы нам предсказывать среднее значение  $y$  с учетом наблюдаемой ее изменчивости.

Во многих ситуациях, однако,  $y$  не является непрерывной величиной, а принимает всего два возможных значения. Обычно единицей в этом случае представляют осуществление какого-либо события (успех), а нулем - отсутствие его реализации (неуспех).

Среднее значение  $y$  - обозначенное через  $p$ , есть доля случаев, в которых  $y$  принимает значение 1. Математически это можно записать как  $p = P(y = 1)$  или  $p = P(\text{"Успех"})$ .

*ROC-кривая* или *кривая ошибок* - показывает зависимость количества верно классифицированных положительных объектов (по оси  $y$ ) от количества неверно классифицированных отрицательных объектов (по оси  $x$ ).

В терминологии ROC - анализа первые называются истинно положительным, вторые – ложно отрицательным множеством. При этом предполагается, что у классификатора имеется некоторый

параметр, варьируя который, мы будем получать то или иное разбиение на два класса. Этот параметр часто называют порогом, или точкой отсечения. В зависимости от него будут получаться различные величины ошибок I и II рода.

В логистической регрессии порог отсечения изменяется от 0 до 1 – это и есть расчетное значение уравнения регрессии. Будем называть его рейтингом.

Введём ещё несколько определений:

*TP (True Positives)* – верно классифицированные положительные примеры (так называемые истинно положительные случаи);

*TN (True Negatives)* – верно классифицированные отрицательные примеры (истинно отрицательные случаи);

*FN (False Negatives)* – положительные примеры, классифицированные как отрицательные (ошибка I рода). Это так называемый «ложный пропуск» – когда интересующее нас событие ошибочно не обнаруживается (ложно отрицательные примеры);

*FP (False Positives)* – отрицательные примеры, классифицированные как положительные (ошибка II рода). Это ложное обнаружение, т.к. при отсутствии события ошибочно выносится решение о его присутствии (ложно положительные случаи).

Что является положительным событием, а что – отрицательным, зависит от конкретной задачи. Например, если мы прогнозируем вероятность наличия заболевания, то положительным исходом будет класс «Больной пациент», отрицательным – «Здоровый пациент». И наоборот, если мы хотим определить вероятность того, что человек здоров, то положительным исходом будет класс «Здоровый пациент», и так далее.

При анализе чаще оперируют не абсолютными показателями, а относительными – долями, выраженными в процентах:

Доля истинно положительных примеров (*True Positives Rate*):

$$TPR = \frac{TP}{TP + FN} \cdot 100 \%$$

Доля ложно положительных примеров (*False Positives Rate*):

$$FPR = \frac{FP}{TN + FP} \cdot 100 \%$$

Введем еще два определения: чувствительность и специфичность модели. Ими определяется объективная ценность любого бинарного классификатора.

*Чувствительность (Sensitivity)* – доля истинно положительных случаев:

$$Se = TPR = \frac{TP}{TP + FN} \cdot 100 \%$$

*Специфичность (Specificity)* – доля истинно отрицательных случаев, которые были правильно идентифицированы моделью:

$$Sp = \frac{TN}{TN + FP} \cdot 100 \%$$

Модель с высокой чувствительностью часто дает истинный результат при наличии положительного исхода (обнаруживает положительные примеры). Наоборот, модель с высокой специфичностью чаще дает истинный результат при наличии отрицательного исхода (обнаруживает отрицательные примеры).

ROC-кривая получается следующим образом:

1. Для каждого значения порога отсечения, которое меняется от 0 до 1 с шагом  $dx$  (например, 0,01), рассчитываются значения чувствительности  $Se$  и специфичности  $Sp$ . В качестве альтернативы порогом может являться каждое последующее значение примера в выборке.

2. Строится график зависимости: по оси  $y$  откладывается чувствительность  $Se$ , по оси  $x$  –  $(100 \% - Sp)$  (сто процентов минус специфичность), или, что то же самое,  $FPR$  – доля ложно положительных случаев.

Численный показатель площади под кривой называется  $AUC$  (*Area Under Curve*). С большими допущениями можно считать, что чем больше показатель  $AUC$ , тем лучшей прогностической силой обладает модель. Однако следует знать, что:

- показатель  $AUC$  предназначен скорее для сравнительного анализа нескольких моделей;

- $AUC$  не содержит никакой информации о чувствительности и специфичности модели.

В литературе иногда приводится следующая экспертная шкала для значений  $AUC$ , по которой можно судить о качестве модели:

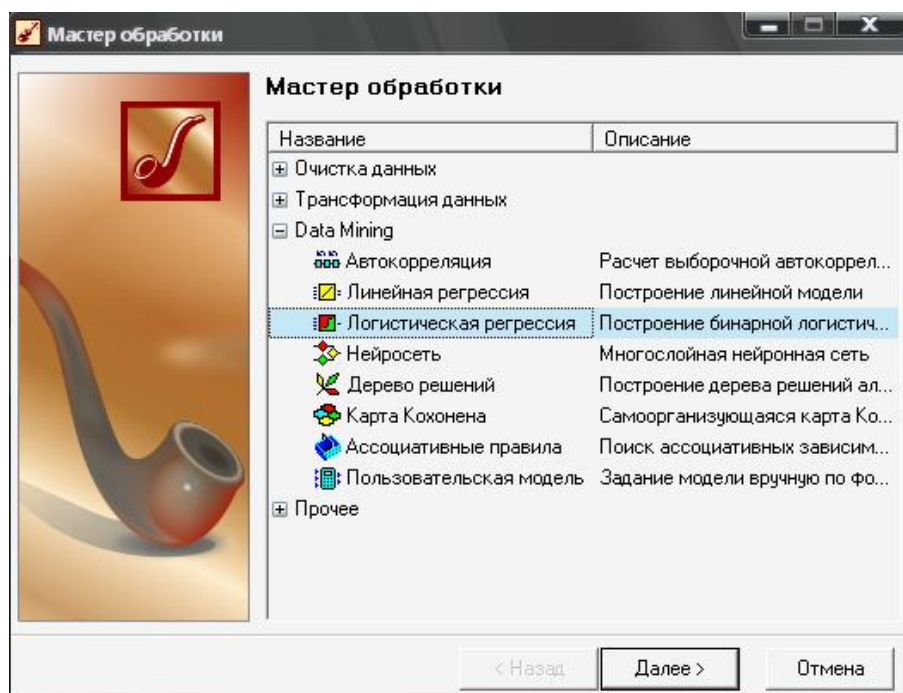
- отличное качество модели – интервал  $AUC$  0,9-1,0;
- очень хорошее качество модели – интервал  $AUC$  0,8-0,9;
- хорошее качество модели – интервал  $AUC$  0,7-0,8;
- среднее качество модели – интервал  $AUC$  0,6-0,7;
- неудовлетворительное качество модели – интервал  $AUC$  0,5-0,6.

Идеальная модель обладает 100 % чувствительностью и специфичностью. Однако на практике добиться этого невозможно, более того, невозможно одновременно повысить и чувствительность, и специфичность модели. Компромисс находится с помощью порога отсечения, т.к. пороговое значение влияет на соотношение  $Se$  и  $Sp$ . Можно говорить о задаче нахождения оптимального порога отсечения.

### 3.3. Практическая часть

Используя *мастер импорта* и файл с данными, например, C:\ProgramFiles\BaseGroup\Deductor\Samples\CreditSample.txt, создайте новый сценарий и импортируйте данные.

В *мастере обработки* выберите способ обработки «Логистическая регрессия».



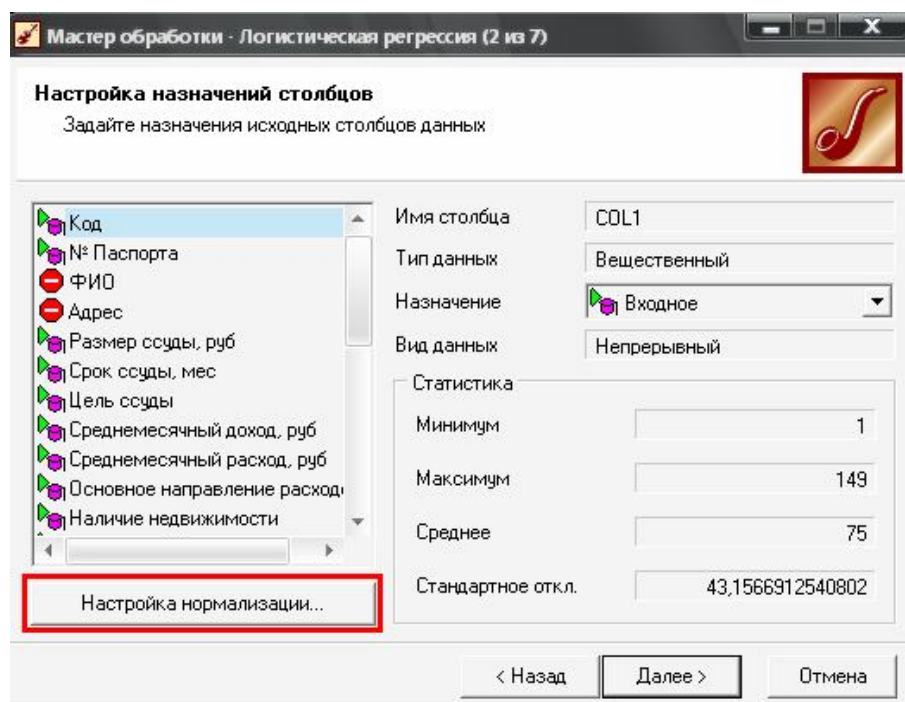
Выбор метода «Логистическая регрессия»

Прежде чем начнется обработка данных, необходимо провести нормализацию полей и настроить обучающую выборку.

*Нормализация полей* проводится с целью преобразования данных к виду, подходящему для обработки средствами АП «Deductor». Например, при построении нейронной сети, линейной модели прогнозирования или самоорганизующихся карт «Входящие» данные

должны иметь числовой тип (т.е. непрерывный характер), а их значения должны быть распределены в определенном диапазоне. В этом случае при нормализации дискретные данные преобразуются в набор непрерывных значений.

Настройка нормализации полей вызывается с помощью кнопки «Настройка нормализации» в нижней левой части окна «Настройка назначения столбцов».

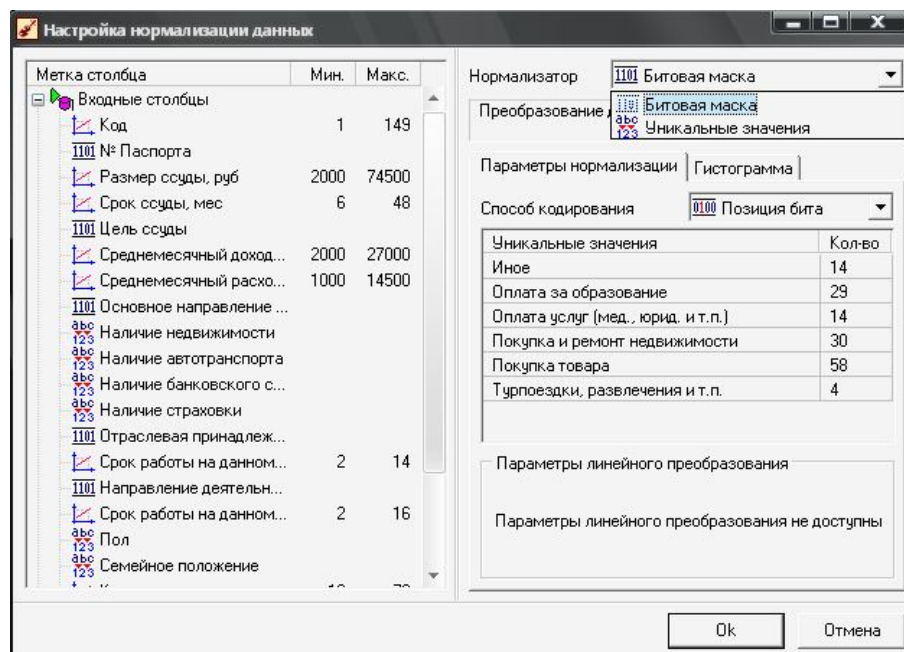


Вызов окна настройки нормализации

В окне «Настройка нормализации данных» слева приведен полный список входных и выходных полей. При этом каждое поле помечено значком, обозначающим вид нормализации:

- *линейная* - линейная нормализация исходных значений;
- *уникальные значения* - преобразование уникальных значений в их индексы;
- *битовая маска* - преобразование дискретных значений в битовую маску.

В правой части окна для выделенного поля отображаются параметры нормализации.



Окно настройки нормализации данных

Для *числовых (непрерывных)* полей с линейной нормализацией дополнительные параметры недоступны. В полях «Минимум» и «Максимум» секции «Диапазон значений» можно посмотреть минимальное и максимальное значения этого поля.

Для *дискретных полей* могут быть использованы два вида нормализации - уникальные значения и битовая маска.

Если дискретные значения преобразуются в битовую маску (т.е. каждому уникальному значению ставится в соответствие уникальная битовая комбинация), то возможны два способа такого преобразования, выбираемые из списка «Способ кодирования»:

1. *Позиция бита* - поле в этом случае представляется в виде  $n$  битов, где  $n$  - число уникальных значений в поле. Каждый бит соответствует одному значению. В 1 устанавливается только бит, соответствующий текущему значению, принимаемому полем, все остальные биты равны 0. Этот способ кодирования используется при малом числе уникальных значений.

2. *Комбинация битов* - каждому уникальному значению соответствует своя комбинация битов в двоичном виде.

*Настройка обучающей выборки* - разбиение обучающей выборки на два множества - обучающее и тестовое - для построения линейной модели.

Мастер обработки - Логистическая регрессия (3 из 7)

**Разбиение исходного набора данных на подмножества**  
 Настройте разбиение исходного множества данных на обучающее и тестовое множества

Способ разделения исходного множества данных: Случайно

Множество	Размер		Порядок сортировки
	В процентах	В строках	
<input checked="" type="checkbox"/> Обучающее	95,00	142	По возрастанию
<input checked="" type="checkbox"/> Тестовое	5,00	7	По возрастанию
<b>ИТОГО:</b>	<b>100,00</b>	<b>149</b>	

Количество строк (всего) 149

< Назад    Далее >    Отмена

Пример настройки обучающей выборки

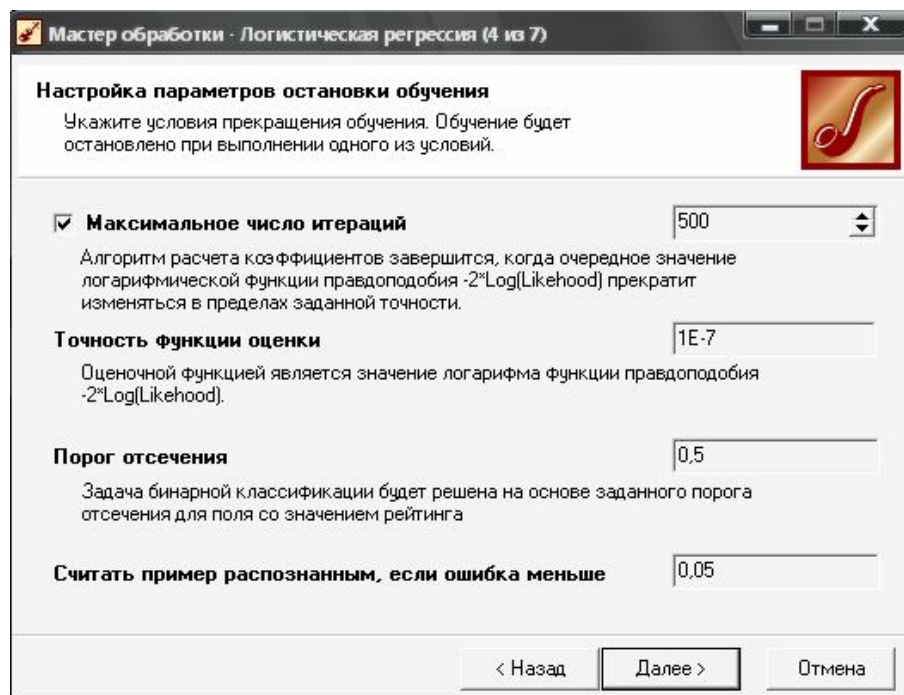
*Обучающее множество* - включает записи, которые будут использоваться в качестве входных данных, а также соответствующие желаемые выходные значения.

*Тестовое множество* - также включает записи, содержащие входные и желаемые выходные значения, но используемое не для обучения модели, а для проверки его результатов.

*Примечание.* Обучение может с большой долей вероятности считаться успешным, если процент распознанных примеров на обучающем и тестовом множествах достаточно велик.

Следующий этап – настройка параметров остановки обучения, которая включает определение максимального числа итераций (заданная точность), задание функции правдоподобия, порога отсечения и допустимость ошибки.





Настройка параметров остановки обучения

Итогом проведения регрессионного анализа будет построенная ROC-кривая.

### 3.4. Задание

1. С помощью мастера импорта откройте файл (например, C:\ProgramFiles\BaseGroup\Deductor\Samples\CreditSample.txt).
2. В *мастере обработки* выберите «Логистическая регрессия».
3. Проведите настройку нормализации полей.
4. Настройте обучающую выборку.
5. Проанализируйте полученные данные.
6. Создайте отчет.



## Лабораторная работа №4. Применение алгоритма кластеризации: самоорганизующиеся карты Кохонена

### 4.1. Основная цель

Научиться использовать метод обработки данных «Самоорганизующиеся карты Кохонена».

### 4.2. Теоретическая часть

Иногда возникают задачи анализа данных, которые с трудом можно представить в математической числовой форме. Это случай, когда нужно извлечь данные, принципы отбора которых заданы нечетко: выделить надежных партнеров, определить перспективный товар и т.п. Таким образом, необходимо на основании имеющихся у нас *априорных* данных получить прогноз на дальнейший период. Для решения этой задачи можно использовать различные методы.

Так, например, наиболее очевидным является применение методов математической статистики. Но тут возникает проблема с количеством данных, ибо статистические методы хорошо работают при большом объеме априорных данных, а у нас может быть ограниченное их количество. При этом статистические методы не могут гарантировать успешный результат.

Другим путем решения данной задачи может быть применение нейронных сетей, которые можно обучить на имеющемся наборе данных. В этом случае в качестве исходной информации используются данные финансовых отчетов различных банков, а в качестве целевого поля – итог их деятельности.

Но при использовании описанных выше методов мы навязываем результат, не пытаясь найти закономерности в исходных данных. Можно попытаться найти эти закономерности с тем, чтобы использовать их в дальнейшем. И тут перед нами возникает вопрос о том, как это сделать.

Существует метод, позволяющий автоматизировать все действия по поиску закономерностей – метод анализа с использованием самоорганизующихся карт Кохонена.

*Самоорганизующаяся карта Кохонена* (англ. *Self-organizing map* — *SOM*) — нейронная сеть с обучением без учителя, выполняющая задачу визуализации и кластеризации. Является методом

проецирования многомерного пространства в пространство с более низкой размерностью (чаще всего двумерное), применяется также для решения задач моделирования, прогнозирования и др.

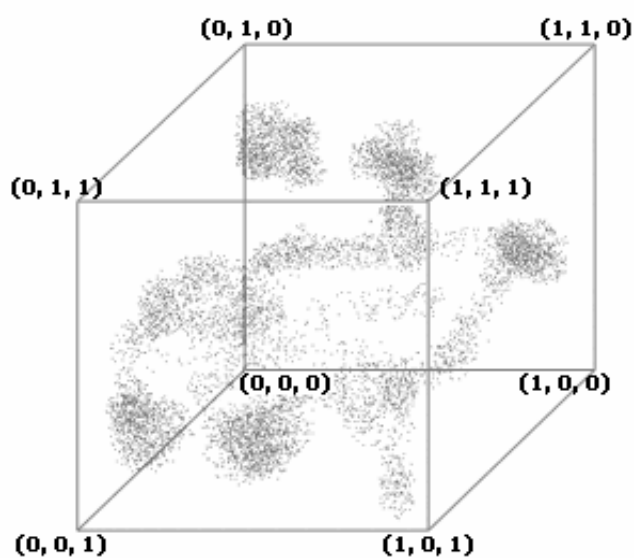
Рассмотрим, как решаются такие задачи и как карты Кохонена находят закономерности в исходных данных. Для общности рассмотрения будем использовать термин «объект» (например, объектом может быть банк, однако описываемая методика без изменений подходит для решения и других задач – например, анализа кредитоспособности клиента, поиска оптимальной стратегии поведения на рынке и т.д.).

Каждый объект характеризуется набором различных *параметров*, которые описывают его состояние. Например, параметрами будут данные из финансовых отчетов. Эти параметры часто имеют числовую форму или могут быть приведены к ней.

Таким образом, нам надо на основании анализа параметров объектов выделить схожие объекты и представить результат в форме, удобной для восприятия.

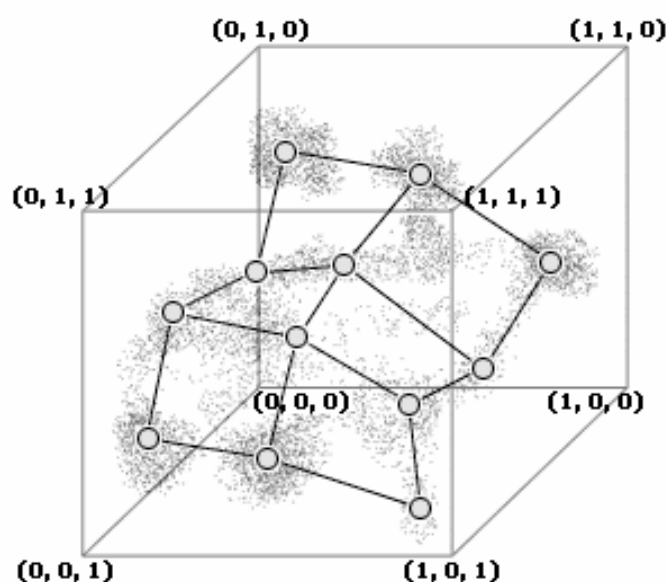
Все эти задачи решаются самоорганизующимися картами Кохонена. Рассмотрим подробнее, как они работают. Для упрощения рассмотрения будем считать, что объекты имеют 3 признака (на самом деле их может быть любое количество).

Теперь представим, что все эти три параметра объектов представляют собой их координаты в трехмерном пространстве (в том самом пространстве, которое окружает нас в повседневной жизни). Тогда каждый объект можно представить в виде точки в данном пространстве, что мы и сделаем (чтобы у нас не было проблем с различным масштабом по осям, пронормируем все эти признаки в интервал  $[0,1]$ ), в результате чего все точки попадут в куб единичного размера. Отобразим эти точки.



Расположение объектов в пространстве

На рисунке мы можем увидеть, как расположены объекты в пространстве, причем легко заметить участки, где объекты группируются, т.е. у них схожи параметры, значит, и сами эти объекты, скорее всего, принадлежат одной группе. Но так легко можно поступить только в случае, когда признаков немного. Значит, нам надо найти способ, которым можно преобразовать данную систему в простую для восприятия, желательно двумерную систему (потому что уже трехмерную картинку невозможно корректно отобразить на плоскости) так, чтобы соседние в искомом пространстве объекты оказались рядом и на полученной картинке. Для этого используем самоорганизующуюся карту Кохонена. В первом приближении ее можно представить в виде «гибкой» сети. Мы, предварительно «скомкав», бросаем сеть в пространство признаков, где у нас уже имеются объекты, и далее поступаем следующим образом: берем один объект (точку в этом пространстве) и находим ближайший к нему узел сети. После этого узел подтягивается к объекту (т.к. сетка «гибкая», то вместе с этим узлом так же, но с меньшей силой подтягиваются и соседние узлы). Затем выбирается другой объект (точка), и процедура повторяется. В результате мы получим карту, расположение узлов которой совпадает с расположением основных скоплений объектов в исходном



Вид пространства после наложения карты

пространстве. Кроме того, полученная карта обладает следующим замечательным свойством – узлы ее расположились таким образом, что объектам, похожим между собой, соответствуют соседние узлы карты. Теперь находим, какие объекты у нас попали в какие узлы карты. Это также определяется ближайшим узлом – объект попадает в тот узел, который находится ближе к нему. В результате данных операций

объекты со схожими параметрами попадут в один узел или в соседние узлы. Таким образом, можно считать, что мы смогли решить задачу поиска похожих объектов и их группировки.

Но на этом возможности карт Кохонена не заканчиваются. Они позволяют также представить полученную информацию в простой и наглядной форме путем нанесения раскраски. Для чего мы раскрашиваем полученную карту (точнее ее узлы) цветами, соответствующими интересующим нас признакам объектов.

Но и это еще не все. Мы можем также получить информацию о зависимостях между параметрами. Нанеся на карту раскраску, соответствующую различным статьям отчетов, можно получить так называемый атлас, хранящий в себе информацию о состоянии рынка. Можно анализировать, сравнивать расположение цветов на раскрасках, порожденных различными параметрами, тем самым получая все новую информацию.

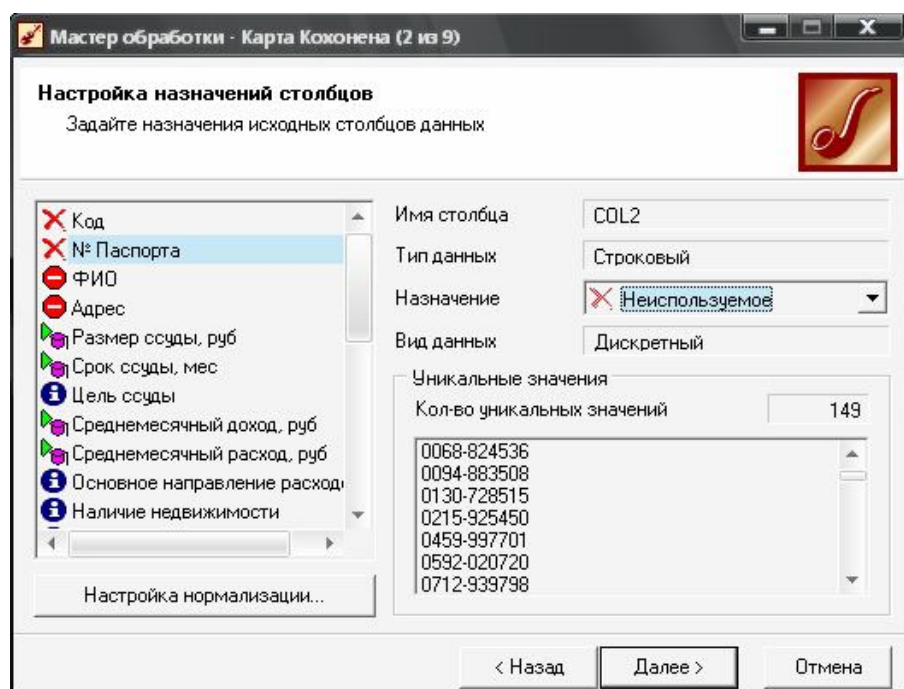
При всем этом описанная технология является универсальным методом анализа. С ее помощью можно анализировать различные стратегии деятельности, производить анализ результатов маркетинговых исследований, проверять кредитоспособность клиентов и т.д.

### 4.3. Практическая часть

Импортируйте в АП «Deductor» исходные данные из файла C:\Program\Files\BaseGroup\Deductor\Samples\CreditSample.txt. Процесс построения карты Кохонена состоит из 10 этапов. Далее рассмотрим эти этапы подробнее.

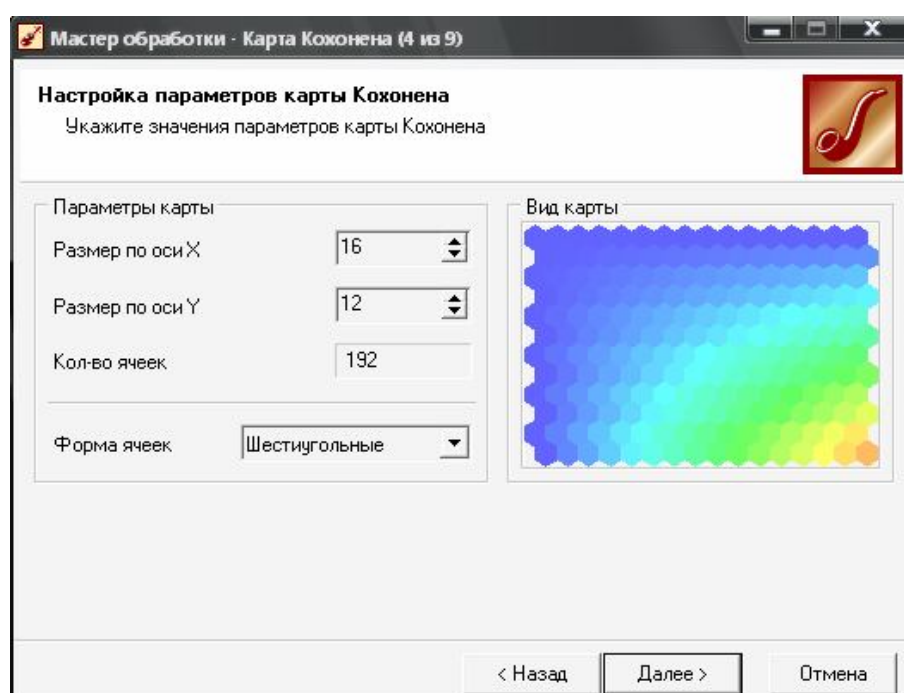
Затем запустите *мастер обработки*, в котором в разделе «Data Mining» выберете способ обработки данных «Карта Кохонена», нажмите «Далее».

В окне настройки назначения столбцов необходимо обозначить столбцы «Код» и «№ паспорта» как «Неиспользуемые» (так как значения этих столбцов уникальны, а это не позволит их классифицировать по общим признакам). Определите поле «Давать кредит» как «Выходное».



Пример настройки назначений столбцов

Настройку обучающей выборки и параметров карты Кохонена можно оставить без изменений.



Настройка параметров карты Кохонена

Настройте параметры остановки обучения, указав уровень допустимой погрешности, если он будет превышен, анализ данного

множества будет прекращен. Можно оставить значения «по умолчанию».

The screenshot shows a window titled "Мастер обработки - Карта Кохонена (5 из 9)". The main heading is "Настройка параметров остановки обучения" (Training Parameters). Below it, a subtitle reads: "Укажите условия прекращения обучения. Обучение будет остановлено при выполнении одного из условий." (Specify the conditions for stopping training. Training will be stopped when one of the conditions is met).

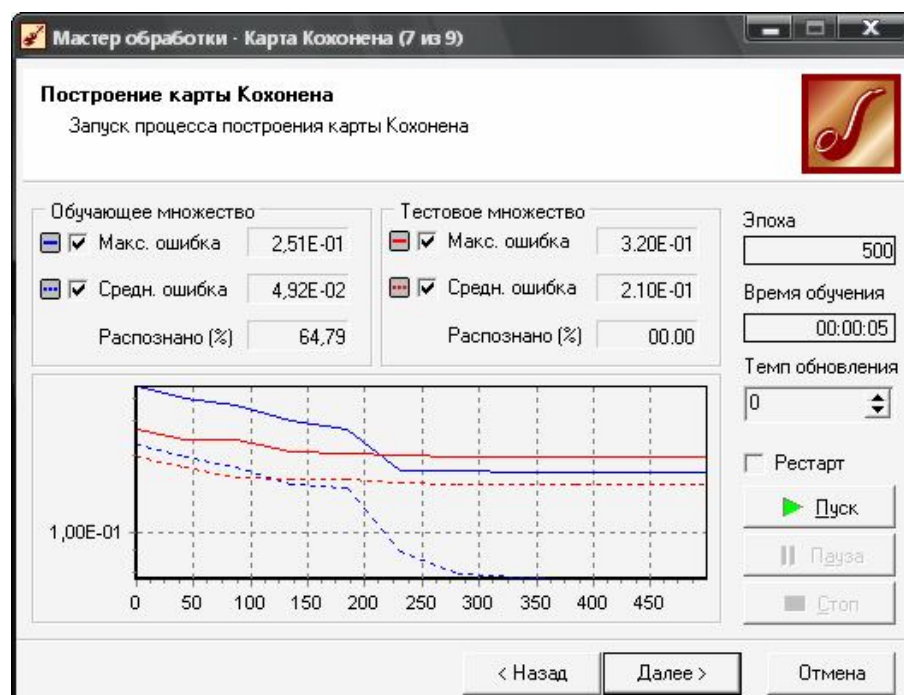
The settings are as follows:

- ☐ Считать пример распознанным, если ошибка меньше: 0,05
- ☒ По достижению эпохи: 500
- Обучающее множество** (Training Set):
  - ☐ Средняя ошибка меньше: [empty field]
  - ☐ Максимальная ошибка меньше: [empty field]
  - ☐ Распознано примеров (%): 0
- Тестовое множество** (Test Set):
  - ☐ Средняя ошибка меньше: [empty field]
  - ☐ Максимальная ошибка меньше: [empty field]
  - ☐ Распознано примеров (%): 0

At the bottom are buttons: "< Назад", "Далее >", and "Отмена".

Настройка параметров остановки обучения

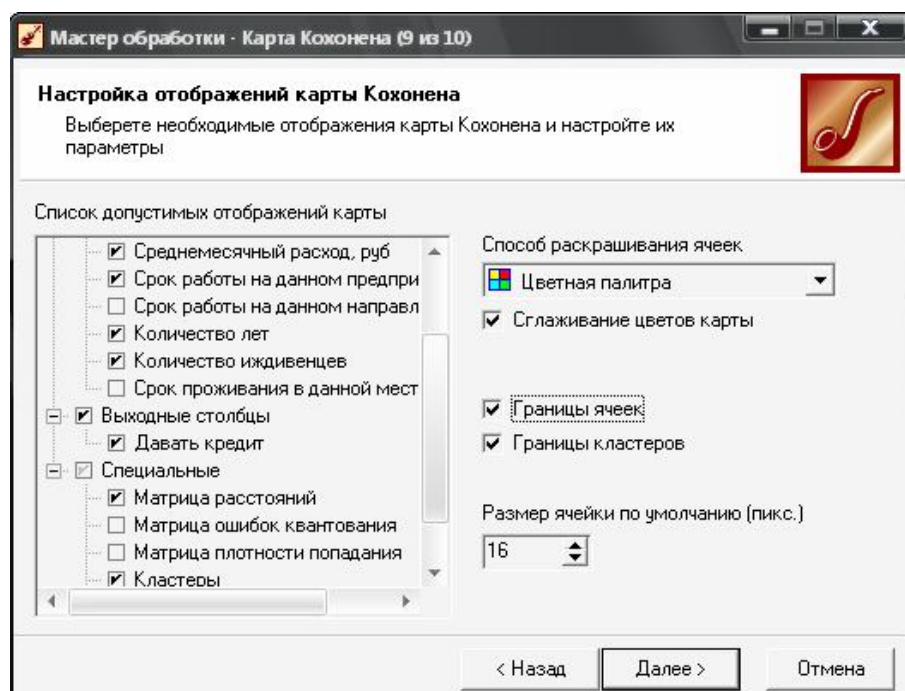
Настройку параметров обучения также оставьте без изменений. Далее запустите процесс построения карты Кохонена, нажав кнопку «Пуск».



Итог построения карты Кохонена

На вкладке «Выбор способа отображения данных» поставьте галочку напротив пункта «Самоорганизующаяся карта Кохонена».

Теперь необходимо провести настройку отображения карты: отметьте разделы «Давать кредит» и «Кластеры» и другие разделы по желанию.

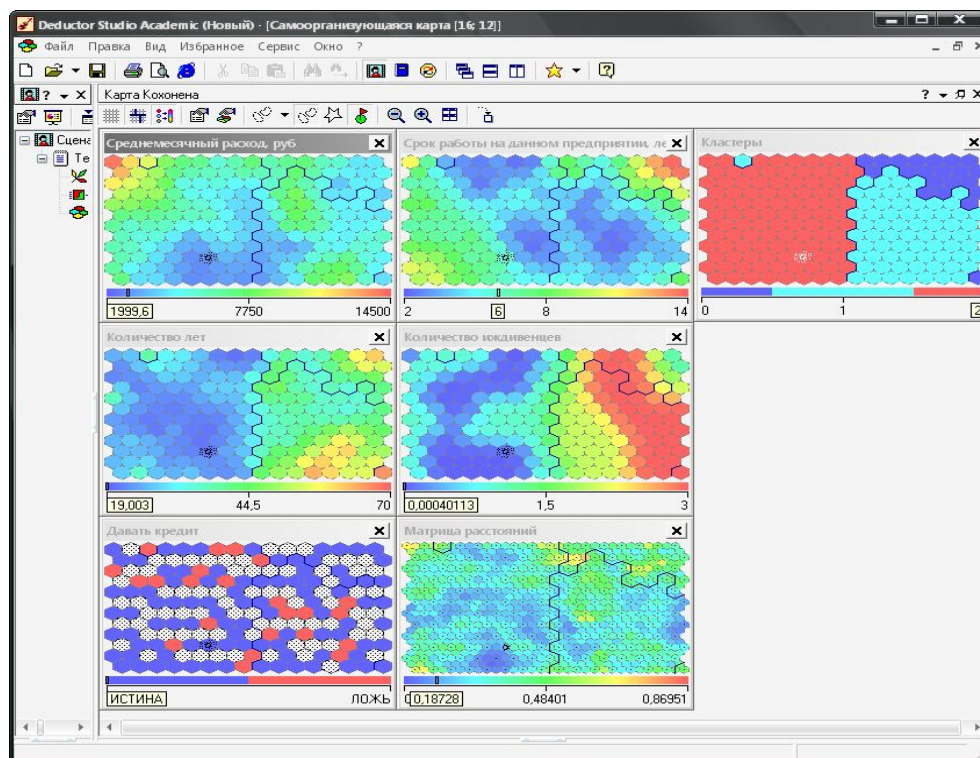


Настройка отображений карты Кохонена

Далее задайте имя, метку и описание карты (по желанию).

В результате получатся карты Кохонена, подобные изображенным на рисунке.





Примеры карт Кохонена

Щелкнув левой клавишей мыши по любому шестиугольнику на любой карте, выделяются соответствующие ему ячейки на остальных картах, в том числе на картах «Давать кредит» и «Кластеры». При этом на шкалах в нижней части карт отобразятся значения соответствующих параметров.

#### 4.4. Задание

1. Выполните описанные выше действия по построению карт Кохонена. Проанализируйте результаты, что можно сказать о вероятности возврата кредита для групп 2, 3 и 4?
2. Используя различные отображения карты Кохонена, постройте 3-4 правила выдачи кредитов.
3. Ответьте на вопросы:
  - для чего используются карты Кохонена?
  - по какому принципу происходит перенос многомерного пространства на пространство меньшей размерности?
4. Подготовьте отчет.



## Лабораторная работа №5. Поиск ассоциативных правил

### 5.1. Основная цель

Научиться выявлять ассоциативные правила с помощью АП «Deductor».

### 5.2. Теоретическая часть

В последнее время неуклонно растет интерес к методам «обнаружения знаний в базах данных». Объемы современных баз данных, которые весьма внушительны, вызвали устойчивый спрос на новые масштабируемые алгоритмы анализа данных. Одним из популярных методов обнаружения знаний стали алгоритмы поиска ассоциативных правил.

Ассоциативные правила позволяют находить закономерности между связанными событиями. Примером такого правила, служит утверждение, что покупатель, приобретающий *хлеб*, приобретет и *молоко* с вероятностью 75 %. Первый алгоритм поиска ассоциативных правил, называвшийся AIS, был разработан в 1993 году сотрудниками исследовательского центра IBM Almaden. На середину 90-х годов прошлого века пришелся пик исследовательских работ в этой области.

#### *Ассоциативные правила (Association Rules)*

Впервые эта задача поиска ассоциативных правил была предложена для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее еще называют анализом рыночной корзины.

Пусть имеется база данных, состоящая из покупательских транзакций. Каждая транзакция - это набор товаров, купленных покупателем за один визит. Такую транзакцию еще называют рыночной корзиной.

Покажем на конкретном примере: «75 % транзакций, содержащих хлеб, также содержат молоко. 3 % от общего числа всех транзакций содержат оба товара». 75 % - это достоверность правила, 3 % - это поддержка, или *хлеб, молоко* с вероятностью 75 %.

Другими словами, целью анализа является установление следующих зависимостей: если в транзакции встретился некоторый набор элементов  $x$ , то на основании этого можно сделать вывод о том, что другой набор элементов  $y$  также должен появиться в этой транзакции. Установление таких зависимостей дает нам возможность находить очень простые и интуитивно понятные правила.

Алгоритмы поиска ассоциативных правил предназначены для нахождения всех правил  $x \Rightarrow y$ , причем поддержка и достоверность этих правил должны быть выше некоторых наперед определенных порогов, называемых соответственно минимальной поддержкой и минимальной достоверностью.

Задача нахождения ассоциативных правил разбивается на две подзадачи:

1. Нахождение всех наборов элементов, которые удовлетворяют порогу минимальной поддержки. Такие наборы элементов называются часто встречающимися.

2. Генерация правил из наборов элементов, найденных согласно п.1 с достоверностью, удовлетворяющей порогу минимальной достоверности.

Один из первых алгоритмов, эффективно решающих подобный класс задач, – это алгоритм APriori. Кроме этого алгоритма в последнее время был разработан ряд других алгоритмов: DHP, Partition, DIC и другие.

Значения для параметров минимальная поддержка и минимальная достоверность выбираются таким образом, чтобы ограничить количество найденных правил. Если поддержка имеет большое значение, то алгоритмы будут находить правила, хорошо известные аналитикам или настолько очевидные, что нет никакого смысла проводить такой анализ. С другой стороны, низкое значение поддержки ведет к генерации огромного количества правил, что, конечно, требует существенных вычислительных ресурсов. Тем не менее большинство интересных правил находится именно при низком значении порога поддержки. Хотя слишком низкое значение поддержки ведет к генерации статистически необоснованных правил.

Поиск ассоциативных правил совсем не тривиальная задача, как может показаться на первый взгляд. Одна из проблем – алгоритмическая сложность при нахождении часто встречающихся наборов элементов, т.к. с ростом числа элементов в  $I$  ( $|I|$ ) экспоненциально растет число потенциальных наборов элементов.

Рассмотрим еще некоторые понятия.

*Транзакция* - множество событий, произошедших одновременно. В нашем случае, каждая транзакция - набор товаров, купленных покупателем за один визит.

*Минимальная и максимальная поддержка.* Ассоциативные правила ищутся только в некотором множестве всех транзакций. Для того чтобы транзакция вошла в это множество, она должна встретиться в исходной выборке количество раз, большее минимальной поддержке и меньше максимальной.

Уменьшение минимальной поддержки приводит к тому, что увеличивается количество потенциально интересных правил, однако это требует существенных вычислительных ресурсов. Одним из ограничений уменьшения порога минимальной поддержки является то, что слишком маленькая поддержка правила делает его статистически необоснованным.

Правило со слишком большой поддержкой с точки зрения статистики представляет собой большую ценность, но с практической точки зрения это, скорее всего, означает то, что, либо правило всем известно, либо товары, присутствующие в нем, являются лидерами продаж, откуда следует их низкая практическая ценность.

Если значение верхнего предела поддержки имеет слишком большое значение, то в правилах основную часть будут составлять товары - лидеры продаж. При таком раскладе не представляется возможным уменьшить минимальный порог поддержки до того значения, при котором могут появляться интересные правила. Причиной тому является просто огромное число правил и, как следствие, нехватка системных ресурсов. Причем получаемые правила процентов на 95 содержат товары - лидеры продаж.

*Минимальная и максимальная достоверность.* Это процентное отношение количества транзакций, содержащих все элементы, которые входят в правило, к количеству транзакций, содержащих элементы, которые входят в условие. Если транзакция - это заказ, а элемент - товар, то достоверность характеризует, насколько часто покупаются товары, входящие в следствие, если заказ содержит товары, вошедшие во всё правило.

Уменьшение порога достоверности также приводит к увеличению количества правил. Значение минимальной достоверности также не должно быть слишком маленьким, так как

ценность правила с достоверностью 5 % настолько мала, что это правило и правилом считать нельзя.

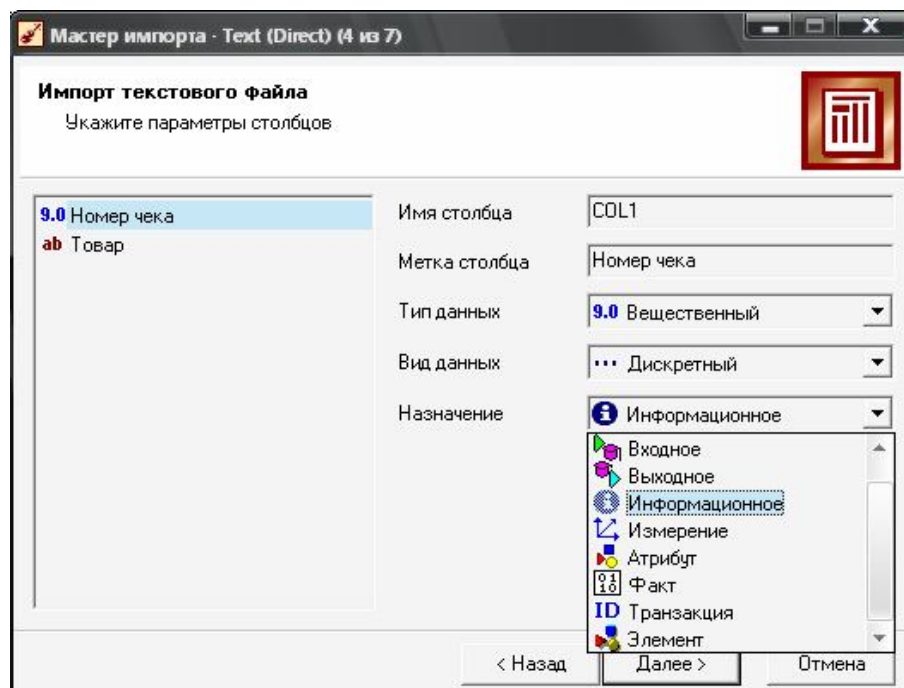
Правило со слишком большой достоверностью практической ценности в контексте решаемой задачи не имеет, т.к. товары, входящие в следствие, покупатель, скорее всего, уже купил.

*Максимальная мощность искомых часто встречающихся множеств.* Если данный параметр указан (флажок установлен), то максимальная мощность (количество элементов) часто встречающихся множеств будет не больше значения этого параметра. Следовательно, любое результирующее правило будет состоять не больше чем из <максимальная мощность> элементов.

Если задать значение параметра максимальная мощность, то можно искать правила, которые состоят не более чем из <максимальная мощность> количества элементов. Например, если нужны только простые правила для оценочного анализа, то значение максимальной мощности следует установить либо в 2, либо в 3. При этом если максимальная мощность равна 2, то все найденные правила будут иметь вид: «Если ТоварI, то ТоварJ». Ограничение поиска часто встречающихся множеств по мощности (количеству элементов в множестве) может также понадобиться, если при указанном значении минимальной поддержки количество часто встречающихся множеств, имеющих большую мощность, слишком велико.

### 5.3. Практическая часть

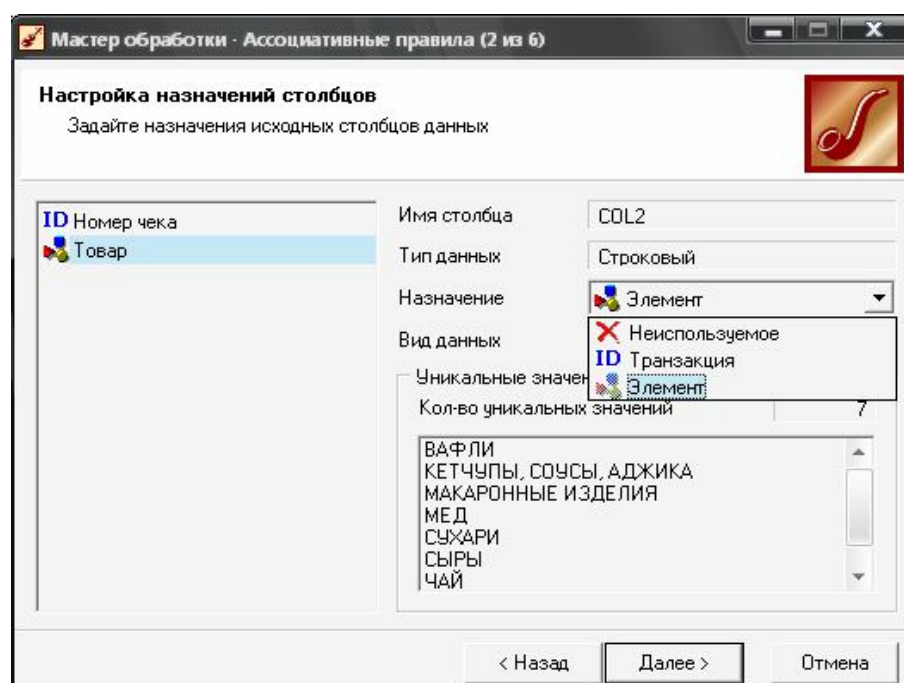
Импортируйте в АП «Deductor» данные файла C:\Program Files\BaseGroup\Deductor\Samples\Supermarket.txt, изменив при этом вид данных столбца «Номер чека» на дискретный.



Изменение параметров импорта данных

В разделе “Data Mining” мастера обработки выберите пункт «Ассоциативные правила».

Установите назначение поля «Номер чека» - транзакция, а поля «Товар» - элемент.



Настройка назначения полей

Установите параметры построения ассоциативных правил, используя информацию, изложенную в теоретической части данного раздела.

Настройка параметров построения ассоциативных правил

Далее запустите процесс поиска ассоциативных правил, нажав кнопку «Пуск».

Выбираем способ отображения данных «Правила» в разделе «Data Mining». В завершение указываем значения полей «Имя», «Метка», «Описание».

#### 5.4. Задание

1. Выполните действия, описанные выше, используя различные параметры построения ассоциативных правил. Сравните полученные результаты, объясните их.

2. Ответьте на вопросы:

- какой товар с наибольшей достоверностью берут с вафлями?
- человек взял *мед* и *сыры*, какой один из товаров он скорее всего не возьмёт?
- назовите 5 самых популярных наборов товаров (в наборе может быть один или несколько товаров).

3. Опишите 4-5 ассоциативных правил, полученных в ходе выполнения работы.

4. Где еще, кроме торговли, можно использовать ассоциативные правила? Приведите примеры.

5. Составьте отчет.

### **3. ИНФОРМАЦИОННО-АНАЛИТИЧЕСКАЯ СИСТЕМА «СЕМАНТИЧЕСКИЙ АРХИВ»**

#### **3.1. Общее описание**

Информационно-аналитическая система «Семантический архив» разработана компанией «Аналитические бизнес решения».

ИАС «Семантический архив» предназначена для автоматизации деятельности аналитических служб коммерческих организаций и государственных структур различного профиля.

Система позволяет организовывать сбор текстовой информации из открытых источников (электронные СМИ, аналитические отчеты экспертов), осуществлять их автоматизированную обработку, эффективное хранение, проведение анализа и генерацию отчетов.

Система предоставляет аналитикам возможность сформировать формальные досье на различные объекты мониторинга – персоны, компании, государственные структуры, а также хранить описания их взаимоотношений и событий, происходящих с ними. Часть отношений и событий могут иметь ссылки на текстовые материалы, в которых они упоминались.

ИАС «Семантический архив» представляет собой программный комплекс, включающий в себя программные компоненты, работающие на сервере и клиентских рабочих местах.

#### **3.2. Возможности системы**

Рассмотрим возможности ИАС «Семантический архив»:

- создание документального архива прессы и внутренних документов компании с развитыми функциями поиска информации, обеспечивающей повышение качества и скорости работы аналитической службы;

- создание архива формальных досье (с развитыми функциями поиска) на персоны и организации, входящие в сферу интересов компании;
- система автоматизирует деятельность операторов по регулярному мониторингу действий персон и компаний по материалам СМИ и другим источникам;
- система автоматизирует деятельность аналитиков при решении аналитических задач (выявление неявных связей между персонами и компаниями, выявление корреляций между происходящими событиями);
- система автоматизирует процесс подготовки отчетов и аналитических записок для руководства.

### **3.3. Технология работы системы**

Поставляемые с системой Интернет-роботы позволяют собирать новости из Интернета, оператор имеет возможность вставлять документы с жесткого диска и импортировать данные из внешних баз данных.

В системе реализовано автоматическое выделение объектов мониторинга из текстов документов и автоматизированное (с участием оператора) выделение событий и отношений между ними.

### **3.4. Типы автоматизированных рабочих мест (АРМ)**

Типы рабочих мест системы:

- конструктор;
- оператор;
- аналитик.

Хранение данных в системе реализовано в объектно-ориентированном виде, что дает аналитикам возможность работать с системой в терминах предметной области.

Уникальным отличием системы от подобных систем является возможность изменения структуры хранилища (добавление новых реквизитов и типовых объектов в процессе эксплуатации системы).

Изменение структуры хранилища может осуществлять сам аналитик без привлечения программистов.

В системе реализован полнотекстовый поиск, поиск по реквизитам документов и по свойствам объектов, отношений и



событий. Аналитик имеет возможность анализировать связи между объектами с помощью семантической сети.

Такой способ визуализации позволяет аналитику увидеть "окружение объекта". Результаты анализа могут быть представлены в виде отчетов различных форматов.

### **3.5. Технологический цикл работы**

Технологический цикл работы включает в себя 5 этапов:

1. *Сбор данных.* Автоматизированный сбор данных из различных источников

2. *Обработка данных.* Автоматизация обработки включает описание пользователем-оператором свойств документа, выделение смысловых конструкций – знаний из текста.

3. *Формирование аналитических запросов.* Формирование информационного среза по документам, объектам и событиям, входящим в область интересов компании. Анализируя параметры найденных объектов, событий или отношений, их связи с «соседними» сущностями, аналитик выявляет интересующие его факты.

4. *Анализ взаимосвязей.* Анализ взаимосвязей объектов путем навигации на семантической сети или автоматический поиск цепочек связей.

5. *Формирование дайджестов и отчетов.* Одной из разновидностей отчета является досье на различных участников исследований или ситуаций в комплексе. Сформированный отчет будет отражать зафиксированный информационный срез модели предметной области.

### **3.6. Технология обработки документов**

1. Описание оператором реквизитов документа (автор, издание, дата публикации и т.д.).

2. Автоматическое выделение из текста объектов мониторинга (персон, компаний, партий и пр.).

3. Автоматизированное выделение (с участием оператора) различных фактов, относящихся к объектам мониторинга, - отношений, состояний и событий.

### **3.7. Технология описания факта**

Описание факта из документа сводится к заполнению полей информационной «карточки». Всего в системе хранится около 300 «шаблонных карточек», описывающих основные типовые отношения и события в экономическом, юридическом, личностном и других аспектах.

Поля карточки могут быть свойством, связью с элементом размерности и связью с объектом.

В системе существует три размерности: «Время», «Географический регион» и «Сфера деятельности».

Общими для всех фактов свойствами являются важность, достоверность и банк данных.

У каждого типа фактов могут быть свои уникальные свойства (сумма контракта, объем производства и т.д.).

В системе регистрируются связи выделяемых фактов с объектами, имеющимися в информационном хранилище. Такая организация хранилища позволяет в дальнейшем визуализировать объекты и факты в виде семантической сети.

### **3.8. Основные функции системы**

К основным функциям системы можно отнести следующие:

- мониторинг новостных сайтов с помощью специализированных Интернет-роботов (поставляемых с системой опционально);
- периодическое импортирование информации из различных реляционных баз данных;
- индексация текстовой и фактографической информации, хранящейся в системе (с целью обеспечения функции быстрого поиска текстовых материалов, объектов мониторинга, отношений и событий);
- полнотекстовый и параметрический поиск;
- визуализация информации в виде таблицы документов, объектов или событий;
- визуализация параметров событий (цена акций компании, количество голосов электората) средствами бизнес-графики;
- визуализация событий и их привязка к карте;

- визуализация объектов, отношений и событий на семантической сети (сетевой вид);
- генерация новостных и аналитических дайджестов, формализованных отчетов и т.д.

### **3.9. Отличительные особенности системы**

Основные отличия от аналогичных систем, представленных на российском рынке:

- организация хранения не только документов, но и объектов мониторинга, событий, данных из внешних баз данных в едином информационном хранилище;
- функция «выделение знаний» - автоматизированное выделение фактов упоминания объектов, отношений и событий из текста документов;
- хранение и представление классификации объектов, отношений и событий в объектно-ориентированном виде (в виде дерева объектов);
- визуализация знаний в виде семантической сети;
- возможность поиска неявных (опосредованных) связей между объектами;
- предоставление профессионального языка объектных запросов (ODL) для аналитиков, который позволяет конструировать запросы произвольной сложности;
- предоставление возможности изменения структуры хранилища в процессе эксплуатации пользователям-аналитикам (без привлечения программистов);
- развитые средства генерации отчетов, позволяющие формировать полноценные отчеты по заранее разработанным шаблонам.

### **3.10. Цели внедрения системы**

Внедрение системы даст следующее:

1. Документальное хранилище, в котором хранятся все важные статьи СМИ и внутренние документы компании. Система обеспечивает к ним мгновенный доступ с помощью функций поиска.
2. Ведение в фактографическом хранилище формальных досье по персонам, компаниям, регионам, крупным проектам.

3. Факты из досье подтверждаются источниками (документами, статьями) в документальном хранилище.

4. Графическое описание ситуации любой сложности. Формирование многоуровневой и многоступенчатой картины развития ситуации.

5. Все отношения между персонами и компаниями (партнеры, конкуренты, заказчики, поставщики и т.п.) наглядно представлены на семантической сети, по которой аналитик осуществляет навигацию от объекта к объекту.

### **3.11. Пользователи системы**

Система предназначена для:

- сотрудников маркетинговых служб коммерческих компаний (например, выявление действий конкурентов);
- сотрудников служб экономической безопасности компании (например, выявление попыток недружественного поглощения компании);
- менеджеров по продажам компаний (например, выявление изменения спроса на товары и услуги, что и как продают конкуренты);
- аналитических служб политических партий (например, как проводят предвыборную кампанию конкуренты);
- PR-служб и др.

## **4. КОМПЛЕКС ЛАБОРАТОРНЫХ РАБОТ ПО ИЗУЧЕНИЮ ИАС «СЕМАНТИЧЕСКИЙ АРХИВ»**

### **Лабораторная работа №1. «Сценарий работы пользователя с модулем поиска «Искатель»**

#### *1.1. Основная цель*

Научиться работать с модулем поиска «Искатель», создавать автоматические задания, изучить язык запросов модуля.

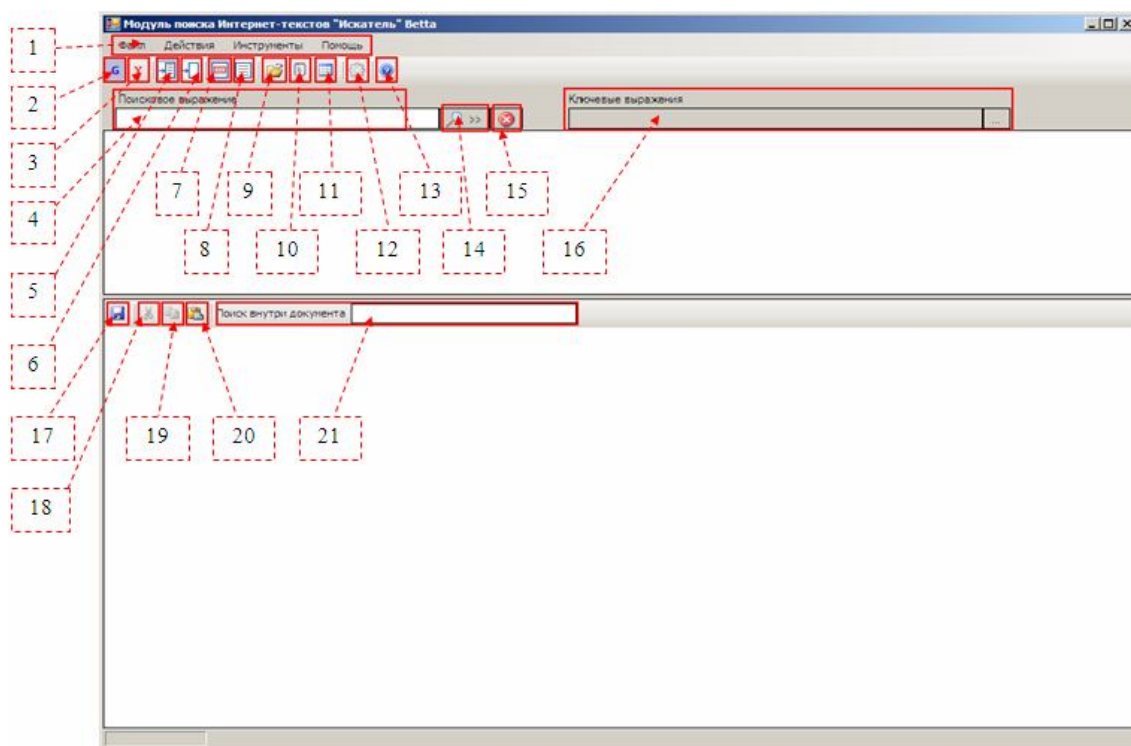
#### *1.2. Пояснения к выполнению работы*

Прежде чем приступить к поиску информации, необходимо определить рамки исследуемой области, поставить конкретные цели исследования и сформулировать вопросы. Естественно, что не важной информации нет, и заранее не известно, в какой из исследуемых областей будет найден ответ. Но поиск во всех областях сразу приведет к накоплению лишней информации, не касающейся исследования. Только заранее установленные рамки помогут наиболее эффективно спланировать процесс поиска.

Рассмотрим основные функции «Семантического архива»:

- автоматический сбор тематической информации из сети Интернет;
- автоматическое создание базы текстовых документов;
- обработка тематических запросов к поисковым сайтам (Yandex, Google и др.).

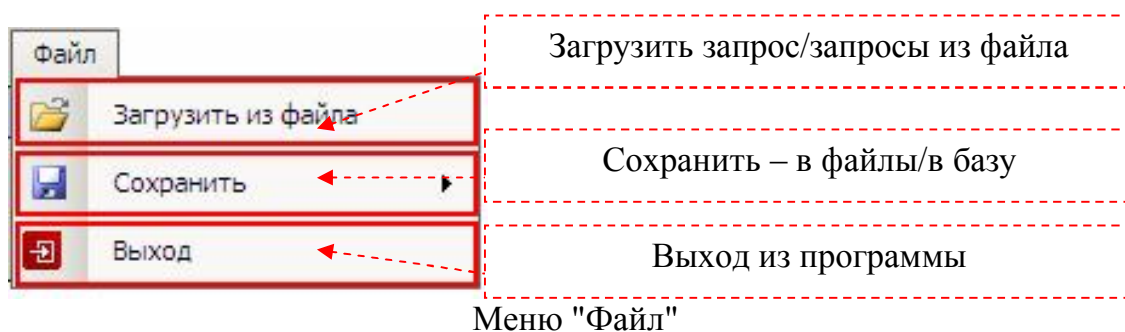
## Элементы окна модуля «Искатель».



Рабочее окно программы «Искатель»:

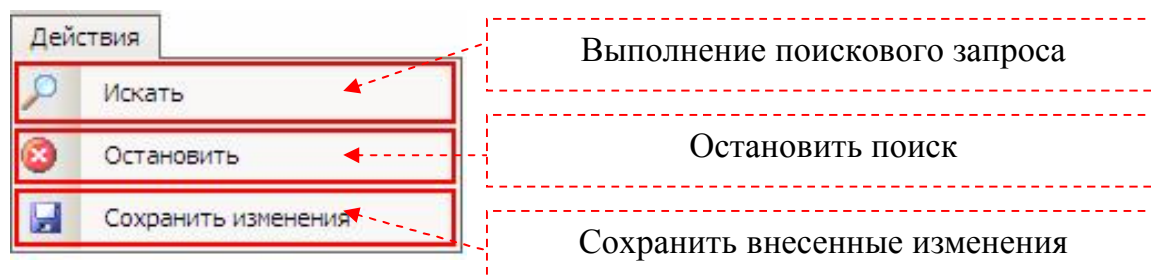
- 1 – меню модуля; 2 – выбор поиска в Google; 3 – выбор поиска в Yandex;  
 4 – строка ввода поискового запроса; 5 – исключать пустые ссылки;  
 6 – показывать пустые ссылки; 7 – показывать фрагмент строки,  
 где найден поисковый запрос; 8 - показывать весь текст Интернет-  
 страницы; 9 – загрузить запрос из файла; 10 – сохранить найденные  
 результаты в файл; 11 - сохранить найденные результаты в базу;  
 12 - настройка модуля; 13 - вызов справки; 14 - выполнить поиск;  
 15 - остановить поиск; 16 - список ключевых выражений; 17 - сохранить  
 внесенные изменения в текст статьи; 18 - вырезать из текста  
 выделенный фрагмент; 19 - копировать выделенный фрагмент  
 текста; 20 - вставить из буфера обмена в текст; 21 - поиск/подсветка  
 выражений в тексте статьи

## Элементы меню «Файл».



Меню "Файл"

### *Меню «Действия».*



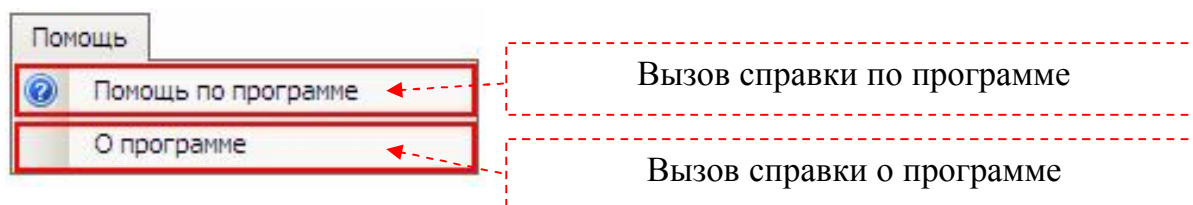
Меню "Действия"

### *Меню «Инструменты».*



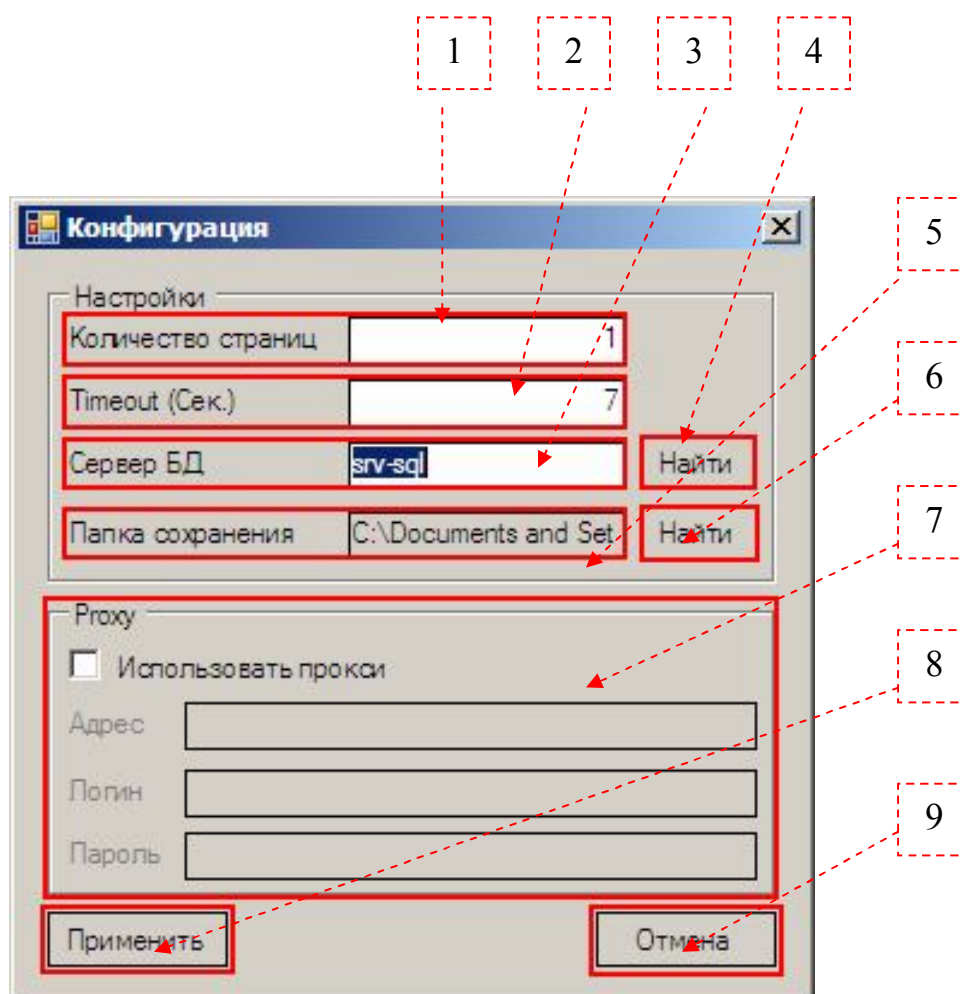
Меню "Инструменты"

### *Меню «Помощь».*



Меню "Помощь"

### Окно настройки конфигурации.



#### Вкладка "Конфигурация":

1 - количество просматриваемых страниц модулем поиска в Интернет-поисковике, по которым будет проводиться поиск; 2 - затрачиваемое время (в секундах) модулем поиска на открытие ссылки; 3 - выбор сервера и базы для сохранения найденной информации; 4 - поиск сервера и базы; 5 - выбор папки для сохранения найденной информации в текстовые файлы; 6 - поиск папки для сохранения найденной информации в текстовые файлы; 7 - настройки прокси-сервера; 8 - применить изменения в настройках; 9 - отменить изменения в настройках

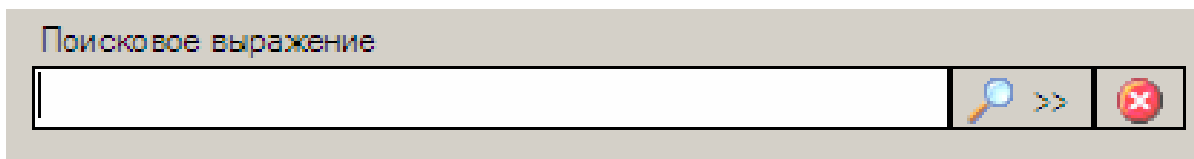
### Поиск по запросу

При поиске по одному запросу необходимо ввести запрос в поле «Поисковое выражение» и нажать клавишу «Enter» или кнопку




«Искать».



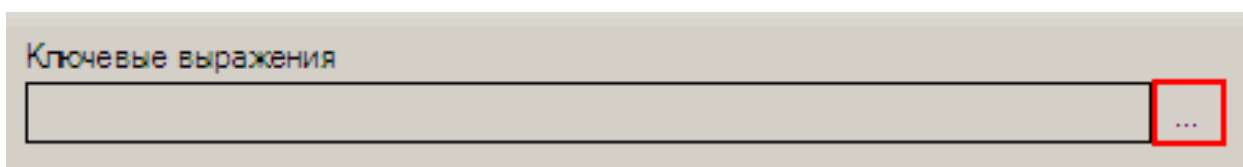
A horizontal input field with the title "Поисковое выражение" (Search expression) in blue. The field is empty. To the right of the field are three buttons: a magnifying glass icon, a ">>" icon, and a red "X" icon.

Поле для ввода поискового выражения

### *Ключевые выражения*

Модуль поиска может производить отбор статей по ключевым выражениям. Если задать ключевые выражения, тогда при поиске по заданному запросу, в случае когда в тексте Интернет-страницы не найдены заданные ключевые выражения, Интернет-страница будет добавляться в список пустых ссылок .

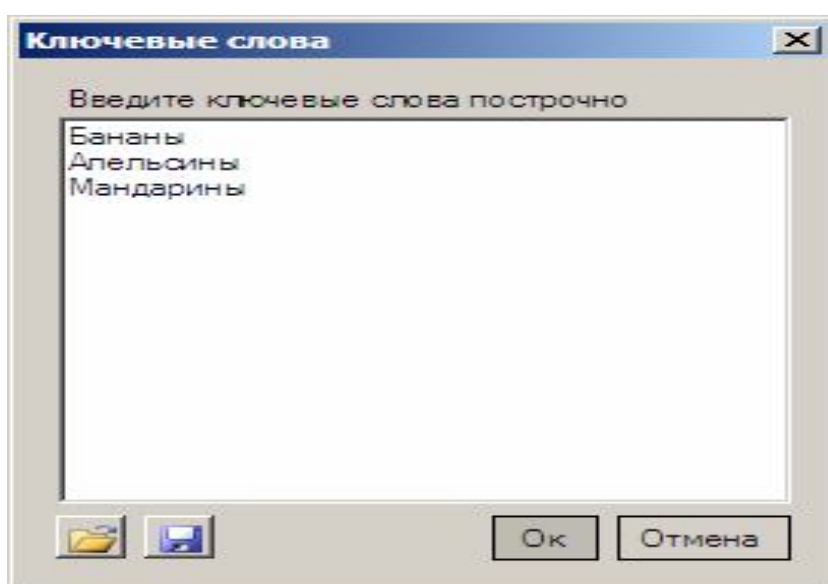
Если ключевые выражения не заданы, поиск ведется без учета ключевых выражений.

A horizontal input field with the title "Ключевые выражения" (Key expressions) in blue. The field is empty. To the right of the field is a small red square button with three dots "..." inside.

Поле для ввода ключевых выражений

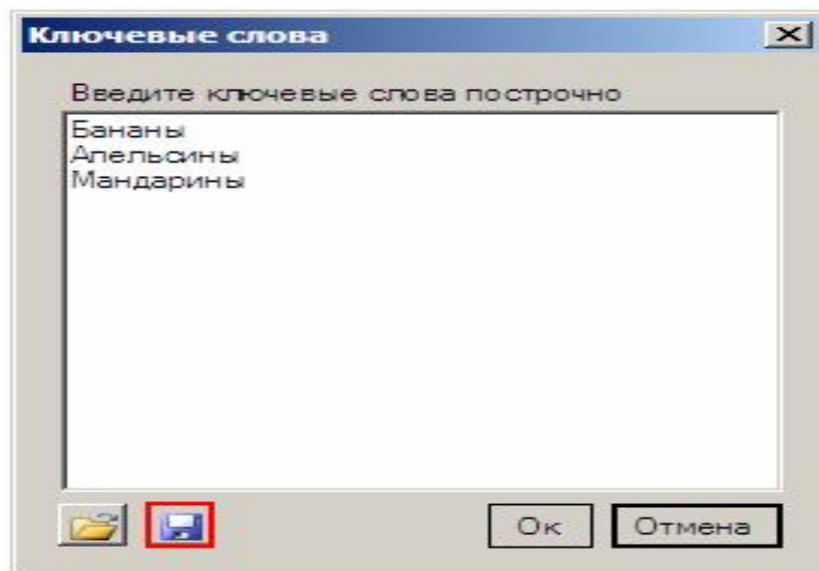
Для ввода ключевых выражений необходимо:

1. Ввести ключевые выражения (ключевые выражения вводятся в столбец).

A dialog box titled "Ключевые слова" (Key words) with a close button "X" in the top right corner. The main text inside says "Введите ключевые слова построчно" (Enter key words line by line). Below this text is a list box containing three items: "Бананы", "Апельсины", and "Мандарины". At the bottom of the dialog box are two buttons: "Ок" (OK) and "Отмена" (Cancel). There are also two small icons on the left: a folder icon and a document icon.

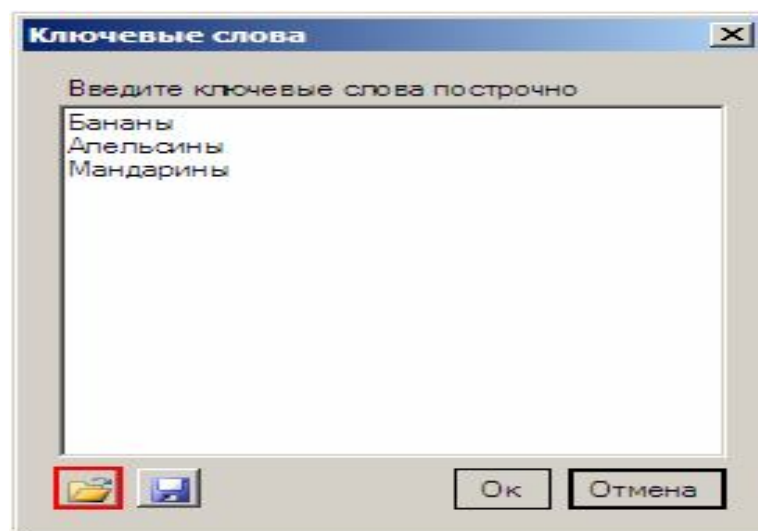
Окно ввода ключевых слов

2. Нажать кнопку «ОК».
3. При желании, если заданные ключевые выражения будут использоваться в дальнейшем, можно их сохранить.



Сохранение ключевых слов

4. Так же можно экспортировать сохраненные в файл ключевые выражения.



Загрузка сохраненных ключевых слов

5. Для загрузки сохраненных файлов-запросов нужно выбрать в меню «Файл» вкладку «Загрузить из файла». В появившемся окне указать путь к файлу-запросу и нажать кнопку «Открыть».

## *Создание задания в Windows для модуля поиска «Искатель»*

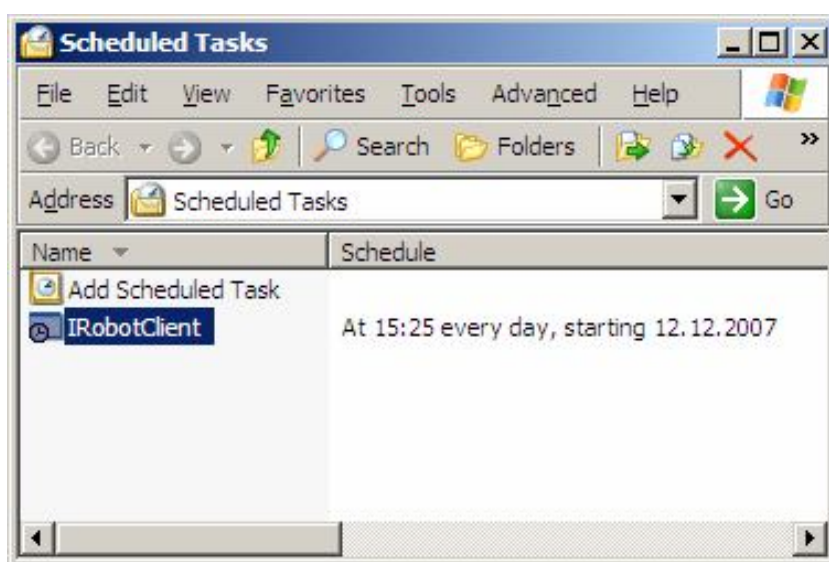
Можно создать задание для модуля поиска «Искатель», которое будет автоматически запускаться и производить поиск по одному или нескольким запросам в заданное время.

Перед созданием задания необходимо ввести в самом модуле поиска «Искатель» настройки сохранения информации, настройки количества страниц поиска и времени, затрачиваемого на страницы и ссылки, по которым будет искать модуль поиска (модуль поиска «Искатель» – меню «Инструменты»).

Если используется прокси-сервер, необходимо его настроить (модуль поиска «Искатель» – меню «Инструменты»).

Пример создания задания:

1. Выбираем в меню «Все программы» - «Стандартные» - «Служебные» - «Назначить задание».



Назначение задания

2. После того как задание создано, его нужно открыть.
3. Далее проводим настройку задания.



Настройка задания

В строке RUN нужно ввести следующие параметры:

"C:\Program Files\ABS\IrobotClient\IRobotClient.exe" (путь к «Искателю») и затем, через пробел "D:\Temp\File1.txt" (созданная папка «Temp» на D диске - можно создать в любом месте, тогда указывается путь к этой папке, а в ней «.txt» файл, в котором сохранены поисковые запросы).

### *Запросы*

При создании запросов можно указать два параметра:

1. *@part* - сохраняется выделенная автоматически часть текста. По умолчанию сохраняется весь текст на странице.
2. *@today* - поиск по текущей дате, реализовано с помощью языка запросов Yandex (date="20071025").

### *Основные операторы языка запросов Yandex*

1. «+» слова из запроса обязательно найдены;
2. «-» исключение слов из результата поиска;

3. «Пробел» все слова входят в предложение, аналог «&»; «|»;
4. «&» логическое «и» в пределах предложения;
5. «~» запрос содержит первое слово, но не содержит второе;
6. «&&» логическое «и» в пределах документа;
7. «'» устойчивые словосочетания;
8. «/» указывается расстояние до слова, например «/2»;
9. «( )» объединение.

### *Специальные запросы*

Title, address, image, url, link (например, '#image='Ленин' выдаст картинку с названием Ленин).

### *1.3. Задание*

1. Определите объект поиска: персона, организация, событие и т.д.
2. Укажите папку для сохранения результатов, настройте количество страниц поиска и времени, затрачиваемого на страницы и ссылки.
3. Создайте текстовый файл с запросом на выбранный вами объект:
  - простой запрос;
  - сложный запрос с использованием ключевых слов (выражений), языка запрос Yandex и «Искатель».
4. Создать автоматическое задание в Windows для поиска в определенное время.

### *Контрольные вопросы*

1. Какие параметры позволяет изменять панель конфигурации в модуле поиска «Искатель»?
2. Какие параметры можно ввести в строке RUN при создании автоматического задания?
3. Перечислите специальные запросы модуля «Искатель» и их назначение.
4. Перечислите основные операторы языка запросов Yandex.

## **Лабораторная работа №2. Добавление данных в базы данных**

### *2.1. Основная цель*

Научиться добавлять текстовые данные в базы данных «Семантического архива» в АРМ «Оператор».

### *2.2. Пояснения к выполнению работы*

Проводя аналитическое исследование, всегда следует вести архив полученных данных, поскольку одни и те же данные спустя время могут толковаться по-разному. Также это необходимо для выявления причинно-следственных связей, которые можно проследить только во временном промежутке.

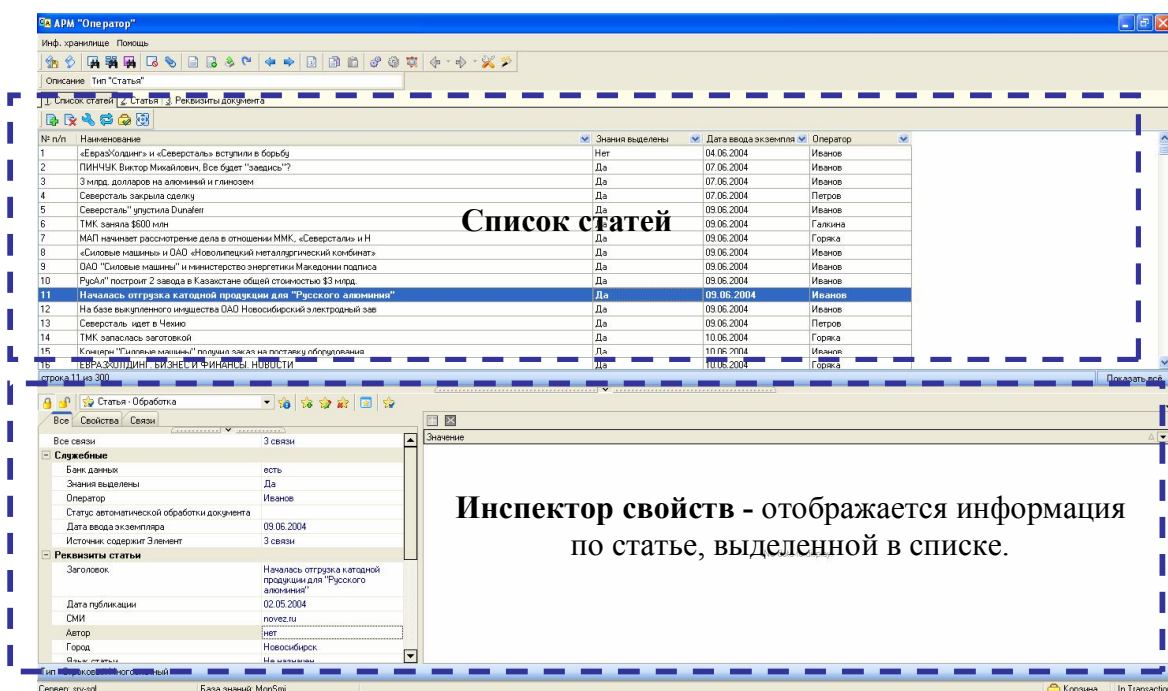
ИАС «Семантический архив» предназначена для решения широкого круга аналитических задач. Для этого в системе создаются базы данных, в которых накапливаются знания об объектах и их действиях. Система поставляется с несколькими готовыми базами данных, которые нуждаются в обновлении, а для работы с новыми объектами возникает необходимость создания архивов путём добавления новых текстовых массивов.

### *Открытие витрины АРМ «Оператор»*

Для открытия АРМ «Оператор» последовательно выберите левой кнопкой мыши пункты меню: «Пуск» - «Все программы» - «Аналитические бизнес решения» - «Семантический Архив» - «АРМ «Оператор».

В диалоговом окне «Настройка подключения» введите имя сервера в поле «Сервер» и выберите значение из списка «База данных». Нажмите кнопку «Продолжить» окна «Настройка подключения».

Витрина АРМ «Оператор» после открытия выглядит следующим образом.



Витрина АРМ "Оператор"

### Добавление данных через «Утилиту добавления документов»

Если вы ищете единичные документы во внешних источниках (Интернете или разных БД, файлах с набором текстов статей или других Windows-приложениях) и при этом нет необходимости хранить информацию в виде текстовых документов (например, новостные статьи), то удобнее всего использовать «Утилиту добавления документов».

Выберите статью для добавления в базу данных «Семантического архива», желательно, чтобы статья содержала: заголовок статьи, дату публикации, источник СМИ, автора статьи.

Чтобы запустить утилиту последовательно, выберите левой кнопкой мыши пункты меню: «Пуск» - «Все программы» - «Аналитические бизнес решения» - «Семантический Архив» - «Утилита добавления документов». После в правом нижнем углу экрана появится значок утилиты.

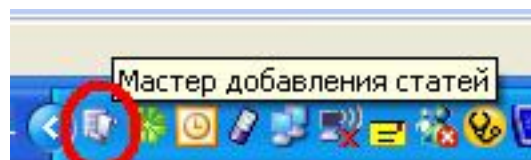
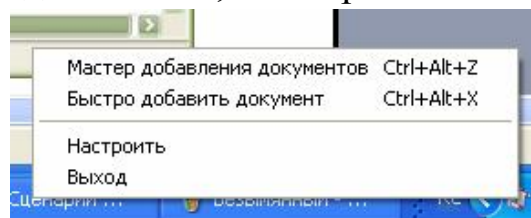
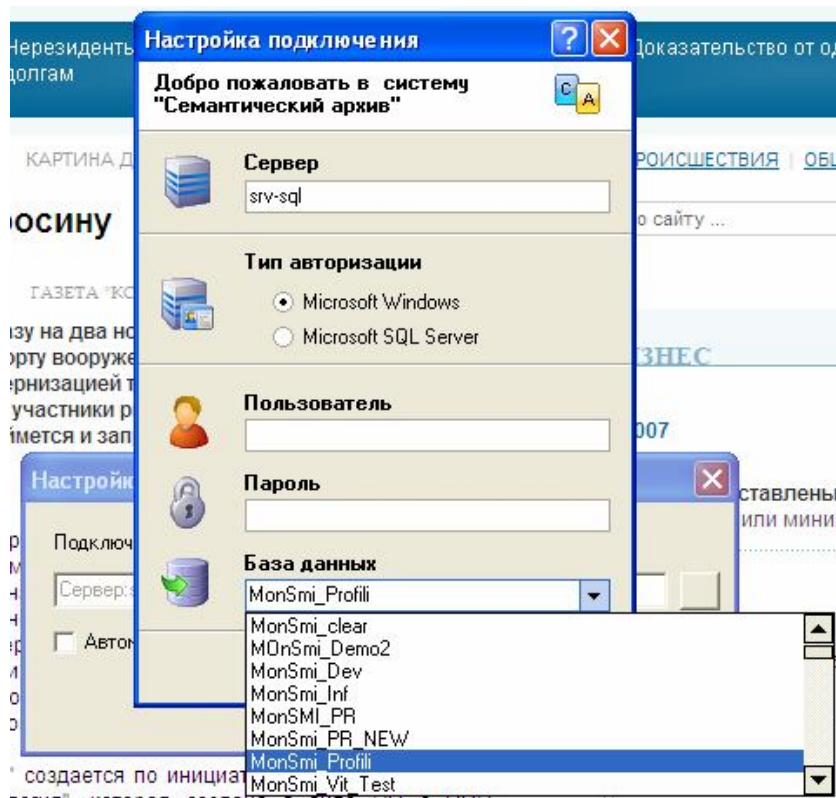


Рис. 4.15. Иконка "Мастер добавления статей"



Меню утилиты добавления документов

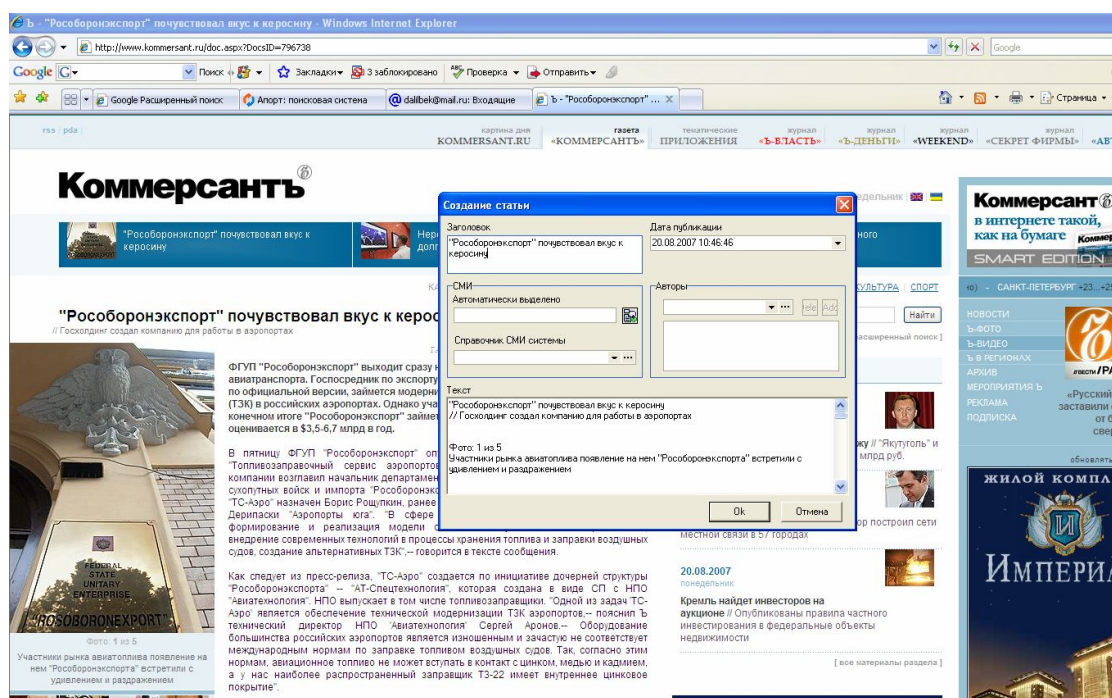
В меню «Утилиты» (меню утилиты вызывается правым нажатием кнопки мышки на значок утилиты), выберете «Настроить» и укажите «Настройка добавления статей в базу». Выбирается та база данных (БД), в которую надо добавить документ.



Выбор БД для добавления документов

Теперь выделите текст статьи вместе с ее реквизитами и копируйте («Ctrl + C» или правая кнопка мышки на выделенном тексте), затем откройте меню утилиты и выберите «Мастер добавления документов» или используйте горячие клавиши «Ctrl + Alt + Z».



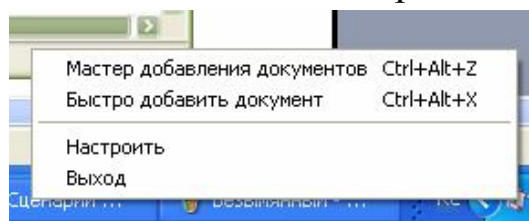


Окно «Мастер добавления документов»

Заголовок статьи автоматически копируется в наименование, дата публикации, источник СМИ и автор статьи также проставляются автоматически (если утилита не проставила автоматически дату, СМИ и автора, можно ввести их вручную), при желании можно откорректировать текст и заголовок статьи.


После нажмие кнопки «ОК» и статья сохранится в базе, которая указана в настройках утилиты.

Для быстрой вставки статьи в «Семантический архив» с помощью утилиты «Добавление документов» необходимо скопировать ее в буфер обмена («Ctrl + C» или правая кнопка мышки и выбрать «Копировать») и в меню утилиты выбирается «Быстро добавить документ» (либо нажатием горячих клавиш «Ctrl + Alt + X»).



Меню утилиты «Добавление документов»

Для вставки в систему одного или нескольких документов через кнопку «Создать документ из файла», сохраните несколько статей в папке в любом текстовом формате, желательно, чтобы статьи содержали: заголовок статьи, дату публикации, источник СМИ, автора статьи.

Для вставки в систему одного или нескольких документов из указанной папки нажмите кнопку  «Создать документ из файла» на главной панели инструментов витрины АРМ «Оператор» или используйте сочетание клавиш «Shift + Ctrl + N».

В открывшемся диалоговом окне выбора документов зайдите в папку, где хранятся файлы статей, и выделите одну или несколько (при помощи сочетания клавиш «Shift + [↑↓]») статей для ввода. Кнопки «Shift + [↑↓]» используются при выделении файлов статей, расположенных в списке друг за другом. При необходимости выделить файлы, находящиеся в разных частях списка, воспользуйтесь сочетанием клавиш «Ctrl + левая клавиша мыши»: удерживая «Ctrl», отметьте левой клавишей мыши нужные вам файлы.

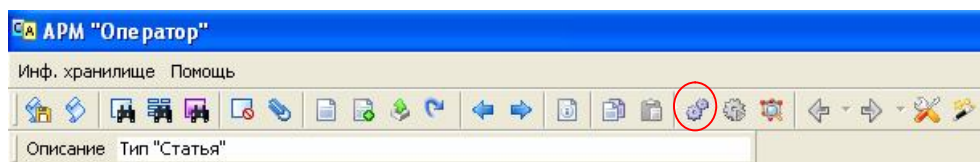
После выбора статей для добавления в базу данных, нажмите кнопку «Открыть», система начнет их обработку.

### *Добавление данных через «Автопапку»*

Если у вас имеется собранная информация, размещенная по тематическим папкам, или вы хотите одновременно сохранять документы в текстовых файлах и автоматически добавлять их в систему, то удобнее воспользоваться «Автопапками». Функцию «Автопапки» можно включать и отключать по мере необходимости. Если документы изначально находятся в разных местах, тогда при помощи средств файловой системы их можно перенести в одну папку, и одновременно с добавлением новых документов уже может идти загрузка. Еще этот способ удобен, если из библиотек поставщиков контента регулярно поступает много статей в какие-то папки файловой системы.

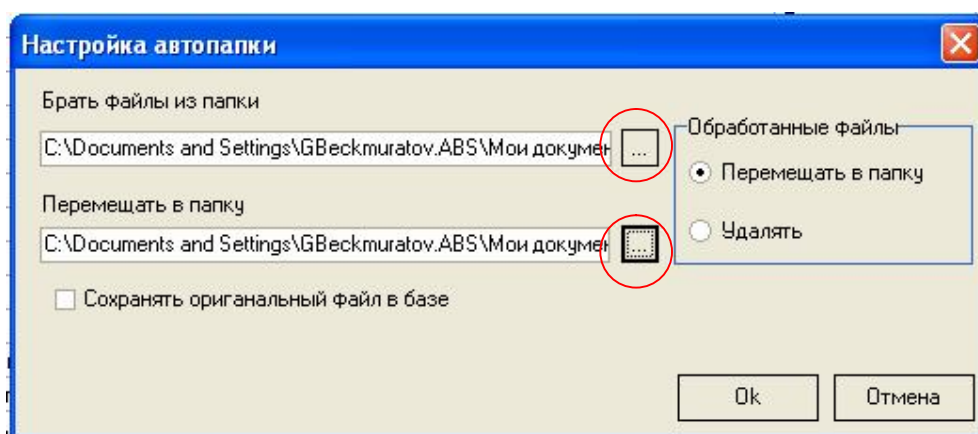
Создайте на своем компьютере две рабочие папки: папку, в которую вы предполагаете помещать новые документы, и папку, в которую система будет перемещать обработанные документы. В первую папку поместите текстовые документы, содержащие данные о некоторых объектах.

Нажмите кнопку  «Настроить автопапку» или используйте сочетание клавиш «Ctrl + Alt + G».




Кнопка «Настроить автопапку» на панели инструментов АРМ «Оператор»

В открывшемся диалоговом окне пропишите адреса папки, из которой система будет брать файлы, и папки, в которую файлы будут перемещаться .




Настройка «Автопапки»

Далее нажмите «ОК».

Для запуска обработки системой статей необходимо нажать кнопку  «Запустить автопапку».

Статьи начнут поступать на обработку в систему и накапливаться в папке, предназначенной для обработанных документов.

### *Примечание*

Необходимо следить за тем, чтобы при изменении путей «Автопапок» кнопка  «Запустить автопапку» не была нажата. Также кнопка должна быть отжата в том случае, если в папке-источнике есть незакрытые файлы офисных приложений (например, MS Word). Иначе в систему продублируются дополнительные файлы, которые создаются офисными приложениями при работе с файлами.

Статьи, добавленные в систему при помощи вставки файлов или «Автопапок», будут иметь незаполненные свойства СМИ, автор, заголовок, поэтому эти свойства необходимо заполнить вручную.

### *2.3. Задание*

1. Найдите биографии любых четырёх известных личностей.
2. Добавьте информацию о первой биографии в БД «Семантического архива» через «Утилиту добавления документов».
3. Добавьте информацию о второй биографии в БД «Семантического архива» через «Утилиту добавления документов» кнопкой «Быстро вставить документ».
4. Добавьте информацию о третьей биографии в БД «Семантического архива» через кнопку «Создать документ из файла».
5. Добавьте информацию о четвёртой биографии в БД «Семантического архива» через «Автопапку».

### *Контрольные вопросы*

1. Перечислите способы добавления данных в БД «Семантического архива».
2. Назовите особенности способов добавления данных в БД «Семантического архива».
3. Какие существуют примечания при добавлении данных в БД «Семантического архива» через «Автопапку» и при помощи вставки?

## **Лабораторная работа №3. Работа в витрине «Сквозного поиска»**

### *3.1. Основная цель*

Поиск выбранных объектов (персоны, организации и т.д.) с заданным значением их свойств и связей из разных БД.

### *3.2. Пояснения к выполнению работы*

Собрав достаточно большой архив данных, необходимо научиться быстрому поиску необходимой информации в собственных архивах. Такая же ситуация возможна при приобретении чужих архивов данных. В таких случаях возникает необходимость появления автоматизированного приложения для сокращения временных затрат в среде структурированных архивов данных.

Сквозной поиск позволяет найти экземпляры с заданным значением их свойств (текстовых или строковых) одного или нескольких искомых типов, хранящиеся в нескольких заданных БД. Для искомого типа можно задать условие на соответствие заданному параметру любого из нескольких выбранных составителем запроса свойств. Поисковый запрос можно настроить таким образом, чтобы он выполнялся по БД, которые находятся на нескольких серверах.

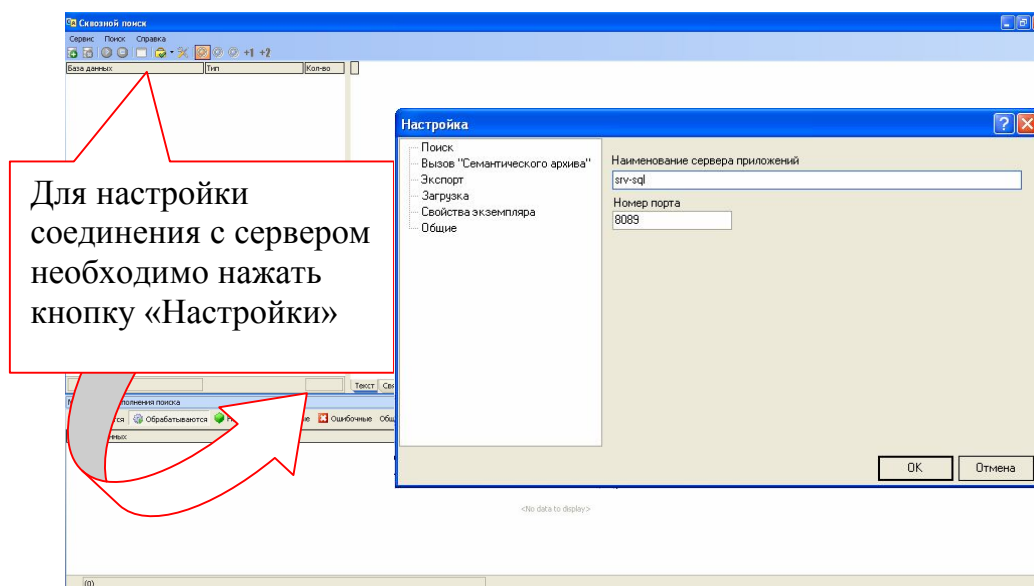
### *Запуск витрины «Сквозной поиск»*

Для запуска витрины «Сквозного поиска» последовательно выберите пункты меню «Пуск» - «Программы» - «Аналитические бизнес решения» - «Семантический архив» - «Сквозной поиск».

При *первоначальном* открытии витрины «Сквозной поиск» система выдаст сообщение о необходимости настройки подключения к серверу с базами данных, в которых будет производиться сквозной поиск.

*Примечание:* Данная операция происходит единожды, лишь при первоначальной загрузке комплекса ИАС «Семантический архив» на компьютер пользователя.

После нажатия кнопки «ОК» откроется витрина утилиты, где необходимо будет настроить *имя сервера*.



Подключение к БД и настройка сервера

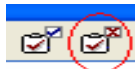
Далее *необходимо* перезапустить витрину «Сквозной поиск»!

## Полнотекстовый поиск

Если раньше для полнотекстового поиска необходимо было заводить объект поиска между кавычками “Иванов”, то теперь данная функция стоит по умолчанию – заводится просто *Иванов*.

## Упрощенное создание запроса для поиска

Упрощенное создание запроса для поиска используется при необходимости ведения поиска по определенным параметрам, например по региону.

Если нужно провести поиск по определенным базам, то необходимо снять выделение с неиспользуемых баз «  ».

Далее, путем фильтрации, выбираем необходимые нам поля.

Drag a column header here to group by that column

Исч.	Сервер	База данных	Тип	Свойство	Параметр	ПТИнде
<input type="checkbox"/>	anbs-sql	[All]	Гибдд по	Дата рождения	Дата рожд	
<input type="checkbox"/>	anbs-sql	64_криминал_саратов_утраченный_паспорт	ГИБДД	Дата рождения	Дата рожд	
<input type="checkbox"/>	anbs-sql	АСУ_Экспресс_2004	ГИБДД	Дата рождения	Дата рожд	
<input type="checkbox"/>	anbs-sql	Акционеры_банков	Автомобиль	Дата рождения	Дата рожд	
<input type="checkbox"/>	anbs-sql	Акционеры_банков_2	Владелец	Дата рождения	Дата рожд	
<input type="checkbox"/>	anbs-sql	БТИ Пермь	ГИБДД Ростов	Дата рождения	Дата рожд	
<input type="checkbox"/>	anbs-sql	БТИ Уфа	АВТО - Москва	Дата рождения	Дата рожд	
<input type="checkbox"/>	anbs-sql	БТИ МО	Автомобиль	Дата рождения	Дата рожд	
<input type="checkbox"/>	anbs-sql	БТИ Омск	Автомобиль	Дата рождения	Дата рожд	
<input type="checkbox"/>	anbs-sql	БТИ_Омск_2004	Автомобиль	Дата рождения	Дата рожд	
<input type="checkbox"/>	anbs-sql	БТИ_Самара_1998	Автомобиль	Дата рождения	Дата рожд	
<input type="checkbox"/>	anbs-sql	БТИ_Хабаровск	ГИБДД	Дата рождения	Дата рожд	
<input type="checkbox"/>	anbs-sql	ВЗД_Справочник	ГИБДД	Дата рождения	Дата рожд	
<input type="checkbox"/>	anbs-sql	ГИБДД Котлас	ГИБДД	Дата рождения	Дата рожд	
<input type="checkbox"/>	anbs-sql	ГИБДД_Уфа_2002	ГИБДД г.Уфа	Дата рождения	Дата рожд	
<input type="checkbox"/>	anbs-sql	ГИБДД_Таганрог	АВТО	Дата рождения	Дата рожд	
<input type="checkbox"/>	anbs-sql	ГИБДД_Тюмень_2002	ГИБДД	Дата рождения	Дата рожд	
<input type="checkbox"/>	anbs-sql	ГИБДД_Москва_2004_09	Автомобиль	Дата рождения	Дата рожд	

Фильтрация по полям

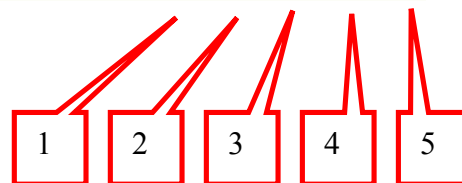
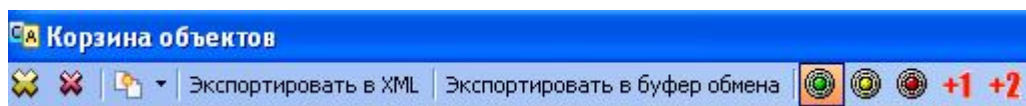
После система покажет поля с выбранными нами параметрами.

Сохраняем данный список баз. При необходимости ведения поиска по данному запросу пользователь просто загружает сохраненный список. Таким методом сохраняются любые параметры для поиска. Также можно загрузить сразу несколько параметров одновременно.

## Визуальное представление результатов поиска

Для более удобного просмотра результатов необходимо поместить их в «Корзину». Система позволяет просмотреть результаты поиска в пяти режимах.



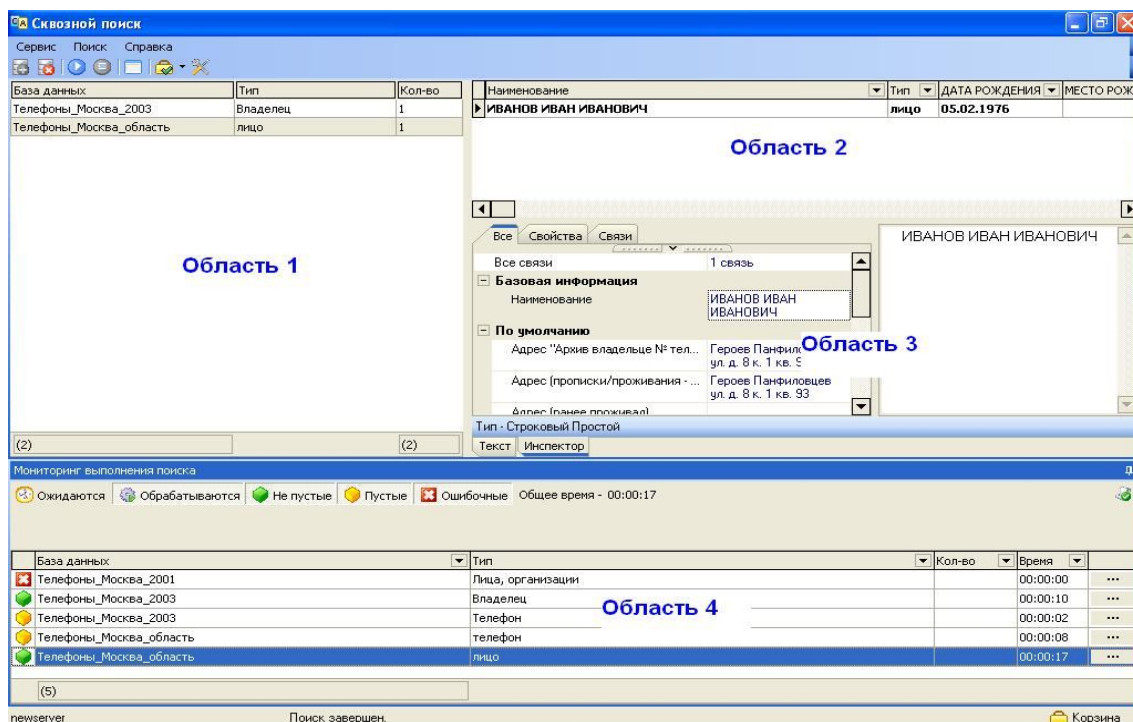


Виды представления информации об объектах:

1 – досье на основной объект; 2 – досье на основной объект, а также список связанных экземпляров; 3 – досье на основной объект, список связанных экземпляров, а также досье на связанные объекты; 4 – досье на основной объект, список связанных объектов, досье на связанные объекты, а также список связанных объектов; 5 – досье на основной объект, список связанных объектов, досье на связанные объекты, список связанных объектов, а также досье на связанные объекты

### *Основные области витрины «Сквозной поиск»*

Окно витрины «Сквозного поиска» состоит из 4-х основных областей.



Окно витрины «Сквозной поиск»

«Область 1» отображает процесс выполнения запроса среди всех заданных типов в базах, по которым осуществляется поиск. Эта

область состоит из 3 колонок. В первых двух колонках указываются имена искомых типов и БД, которой принадлежат эти типы. Третья колонка заполняется только после того, как поиск по данному типу в данной базе завершен. В ней отображается количество найденных экземпляров соответствующего типа, если оно не превышает предела, установленного в окне настройки. Если же запросу удовлетворяет большее число экземпляров данного типа, то в третьей колонке отображается число из окна настроек.

*«Область 2»* отображает найденные объекты в базах.

*«Область 3»* показывает свойства найденных объектов.

*Примечание:* Просмотреть результаты поиска можно как в «Инспекторе свойств», так и строчном режиме.

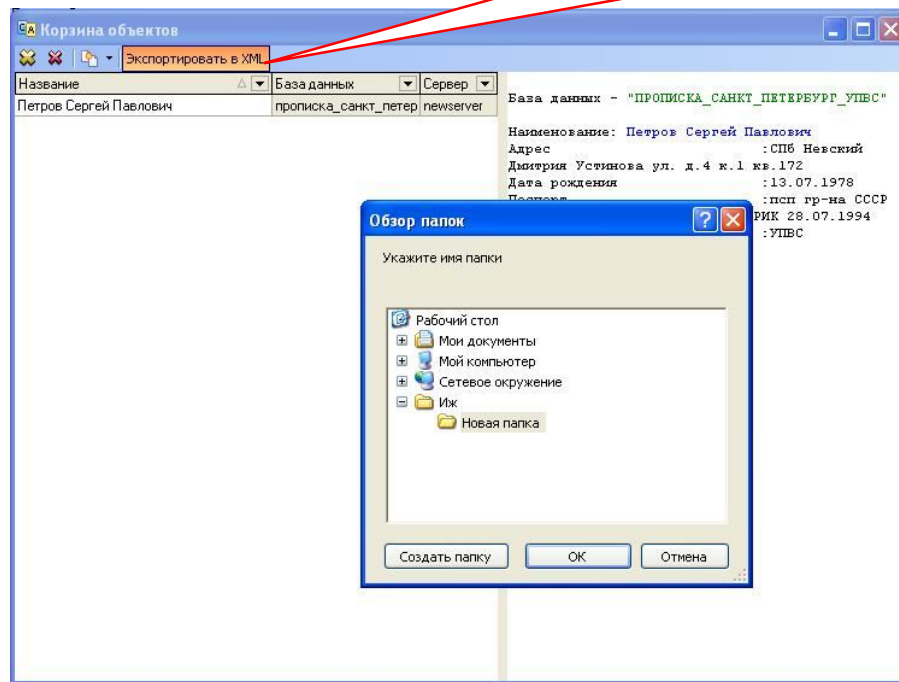
*«Область 4»* отображает полный список доступных баз, по которым производится поиск данных. Данный список может быть отсортирован, т.е. в этой области будут отображаться выбранные базы. Для этого необходимо активировать нужные пункты путем нажатия соответствующих кнопок.

### *Сохранение найденных объектов в виде XML-файла*

Также найденные объекты можно просмотреть в АРМ «Аналитик». Для этого необходимо сохранить данные объекты в виде XML-файла.




Выделяется объект, и нажимаем  
«Экспортировать в XML»



Сохранение в формате XML

В открывшемся окне создается папка, куда помещается созданный XML-файл.

По кнопке  «Запустить Семантический архив» можно запустить витрину АРМ «Аналитик».

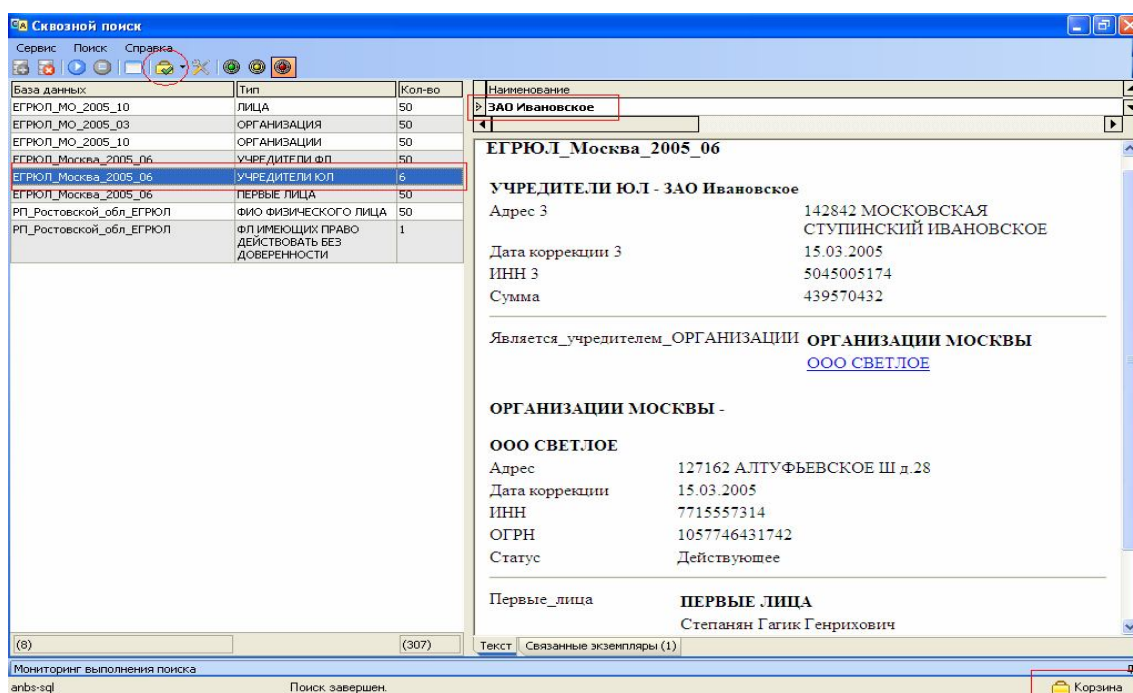
Для просмотра результатов поиска в витрине АРМ «Аналитик» система автоматически подключится к тому серверу, на котором идет поиск, и расположенной на нем БД.

При необходимости АРМ «Аналитик» можно подключить к рабочему серверу и необходимой БД.

Далее, создав новый объект, привязываем к нему результаты поиска, т.е. сохраненный XML-файл.

*Возможность создания нового объекта/факта из текстового содержимого, найденного в «Сквозном поиске» объекта на семантической сети*

После нахождения объекта поиска выделяем объект и помещаем его в «Корзину» со связанными экземплярами.

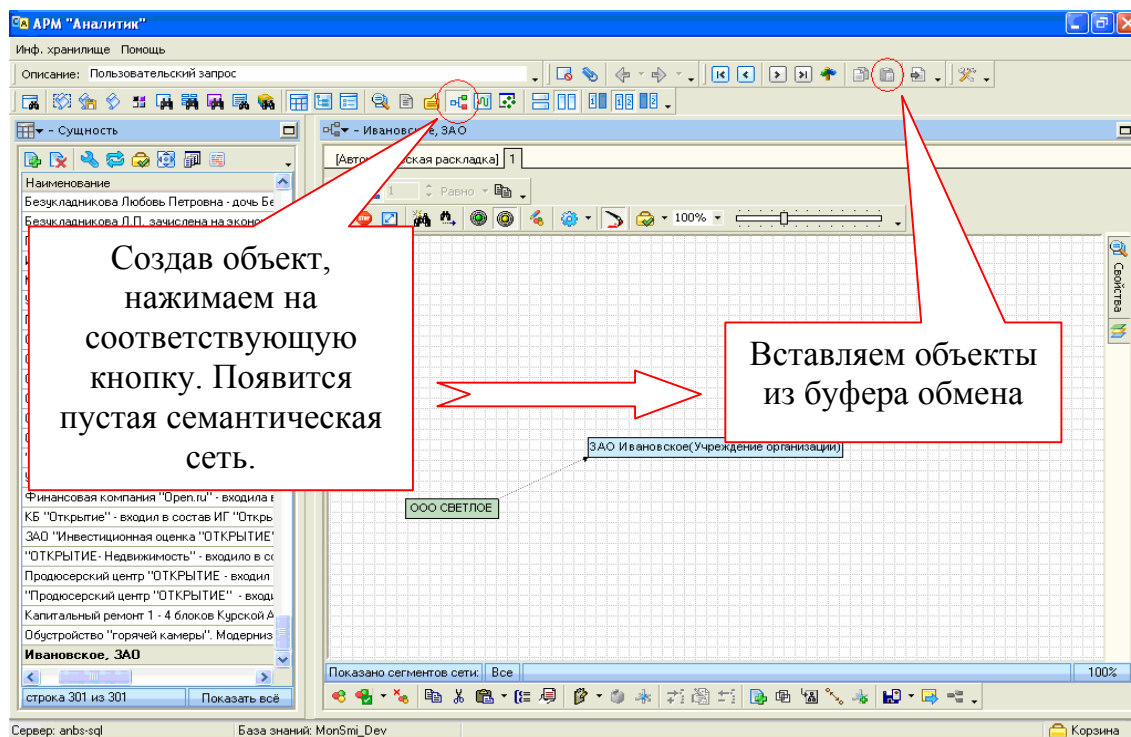


### Перенос объекта в «Корзину»

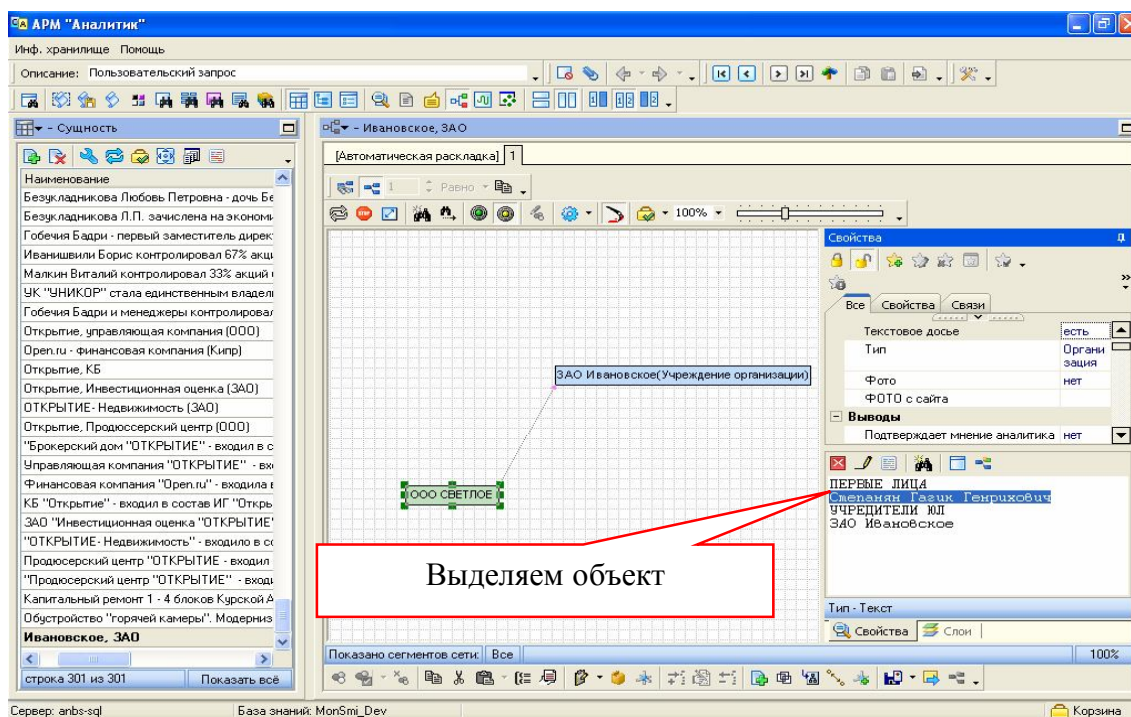
Далее выделяем экземпляры и помещаем их в буфер обмена. Открываем «Корзину» в витрине АРМ «Аналитик». Из буфера обмена в АРМ «Аналитик» добавляем объекты в «Корзину».

На витрине АРМ «Аналитик» необходимо создать определенный объект, создать от этого объекта семантическую сеть (в данном случае она будет пустой).

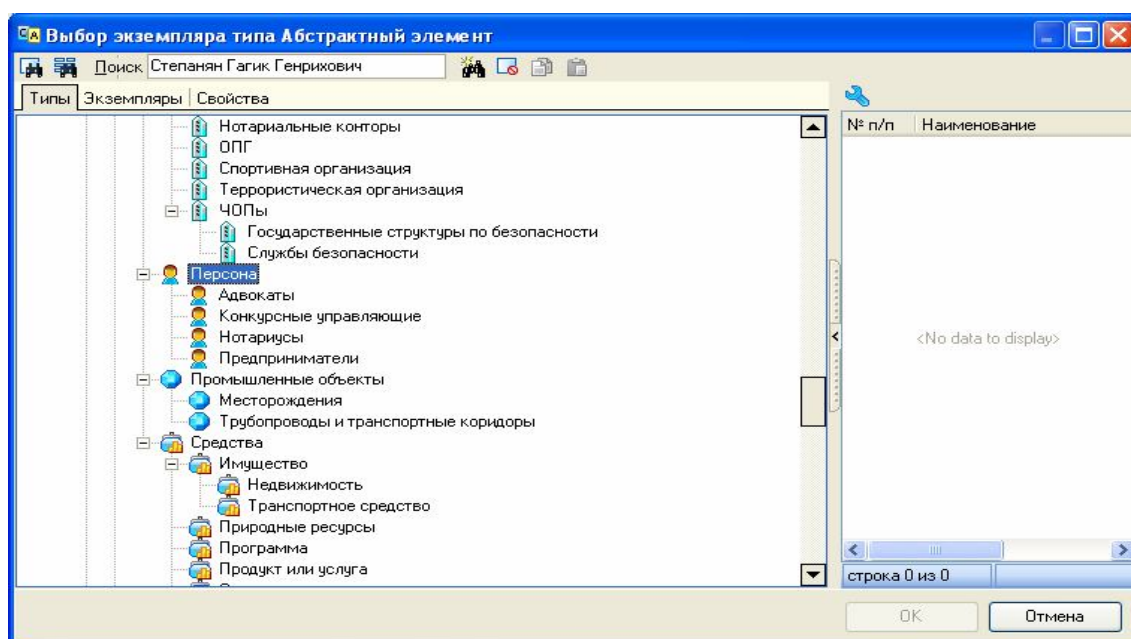
*Примечание:* Это упрощенный метод создания семантических сетей, показывающий лишь принцип вставки.



При просмотре текстового досье объекта выявляем там упоминаемый объект. Далее появится окно создания нового объекта. В нашем случае это персона – Степанян Гарик Генрихович.

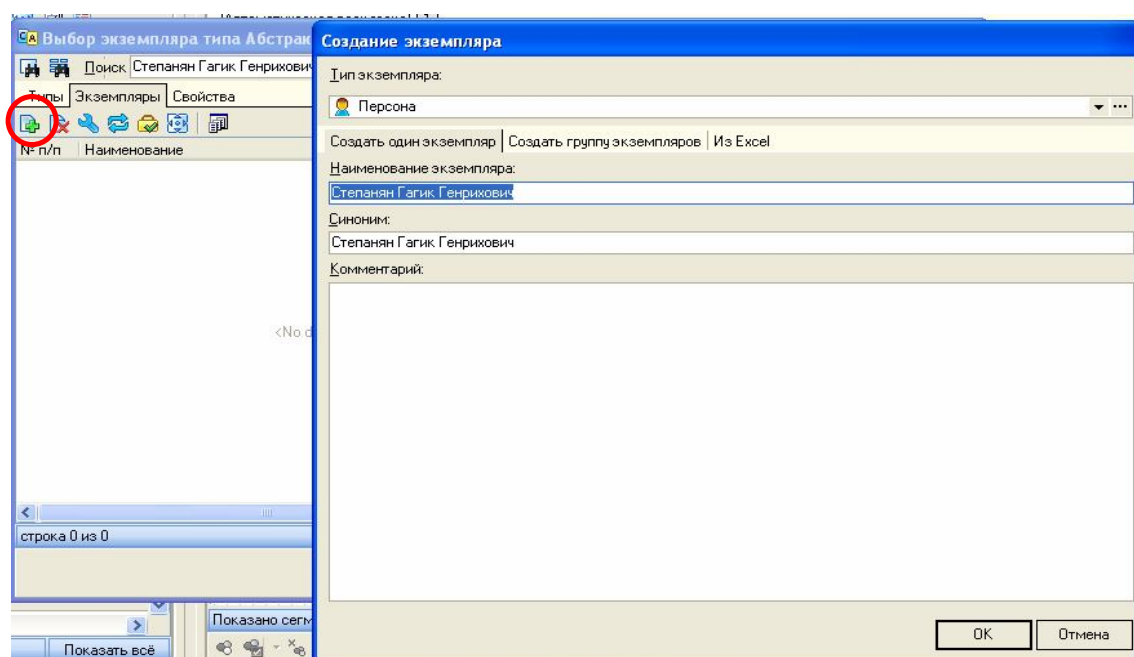


Выбираем нужный «Тип». В нашем примере это «Персона».



Выбор экземпляра нового элемента

Нажимаем на создание «Нового объекта».



Создание экземпляра

Появится окно «Создание экземпляра», где мы можем конкретизировать связь создаваемого объекта с объектом на семантической сети, из текстового содержимого которого был выявлен новый объект.

#### *4.3.3. Задание*

1. Запустите витрину «Сквозной поиск».
2. Найдите информацию о персоне при известных ФИО, годе рождения. Для этого:
  - откройте окно «Запросы»;
  - в окне «Запросы» щелчком левой кнопки мыши выделите название группы баз «Персона по всем полям»;
  - заполните известные поля поиска на закладке «Параметры»;
  - на закладке «Детали запроса» просмотрите параметры БД, по которым будет производиться сквозной поиск;
  - укажите параметры, содержащиеся в тех или иных БД;
  - запустите запрос на выполнение;
  - перенесите данные в «Корзину» и поместите в буфер обмена;
  - экспортируйте данные в файл формата XML.

#### *Контрольные вопросы*

1. Перечислите виды визуального представления результатов поиска.
2. Перечислите основные области витрины сквозного поиска.
3. Перечислите этапы создания нового объекта/факта из текстового содержимого, найденного в «Сквозном поиске» объекта на семантической сети.

## **Лабораторная работа №4. Перенос данных из АРМ «Оператор» в «Аналитик»**

### *4.1. Основная цель*

Научиться переносить обработанные данные из АРМ «Оператор» в АРМ «Аналитик».

### *4.2. Пояснения к выполнению работы*

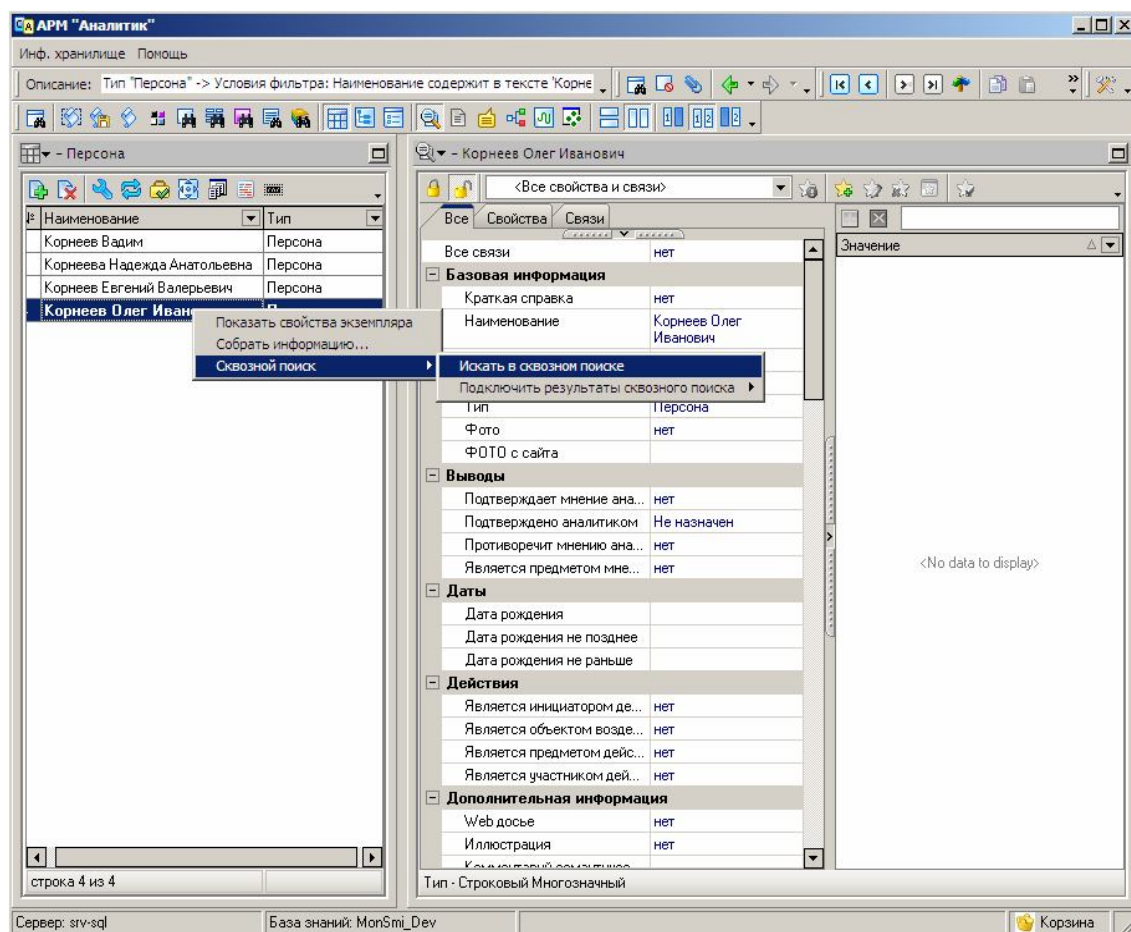
Даже хорошо структурированные данные без визуального отображения не предоставляют возможности легко оценивать картину исследования. Только визуально протянув связи между объектами через их действия или действия над ними, можно построить целостную картину, позволяющую дать наиболее точную оценку. После установления прямых связей объектов можно найти косвенные связи, например объектов через третий объект или действие.

Входными данными для работы АРМ «Аналитик» являются текстовые массивы, переработанные (выделены объекты, их связи и пр.) в АРМ «Оператор». Первоначальным этапом работы аналитика является получение данных об объектах и их связях от оператора, для этого в ИАС «Семантический архив» имеются приспособления «Сквозной поиск» - для нахождения объектов, связей, статей в БД и «Корзина» - для передачи данных.

### *4.3. Поиск объектов и помещение в «Корзину»*

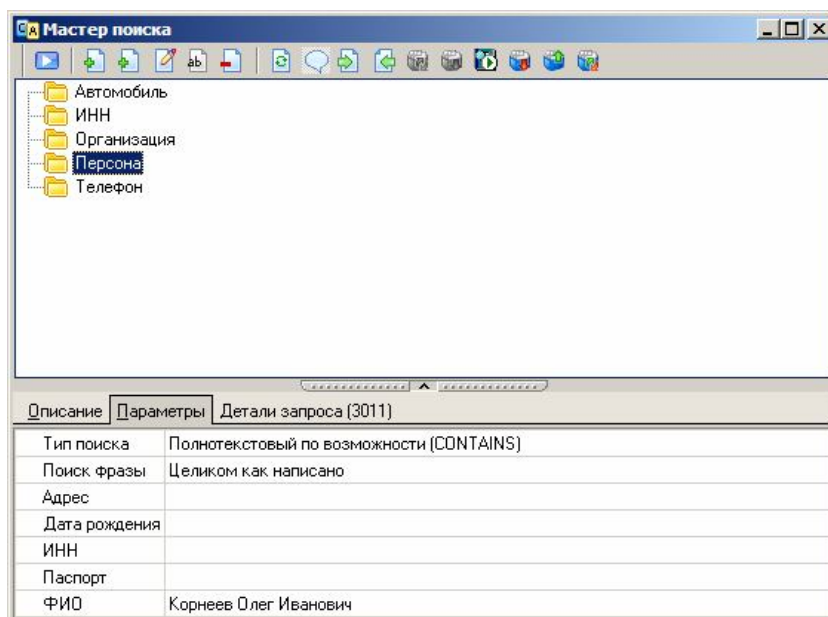
Для поиска объекта в «Сквозной поиск» созданного в АРМ «Аналитике» и привязывания к нему найденных результатов нажмите правой кнопкой мыши на объекте поиска, в выпавшем меню выберите «Сквозной поиск» - «Искать в сквозном поиске».






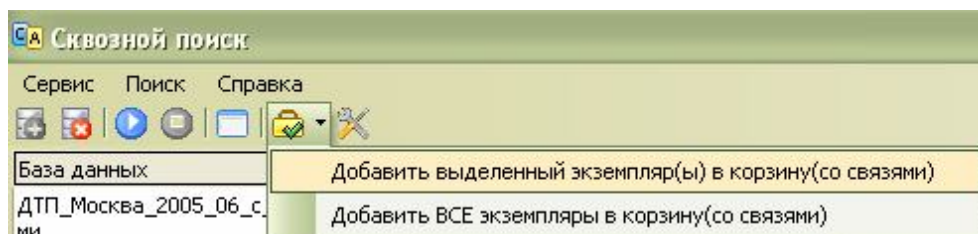
Запуск «Сквозной поиск»

Выберете пункт, по которому будет производиться поиск в окне «Мастер поиска».




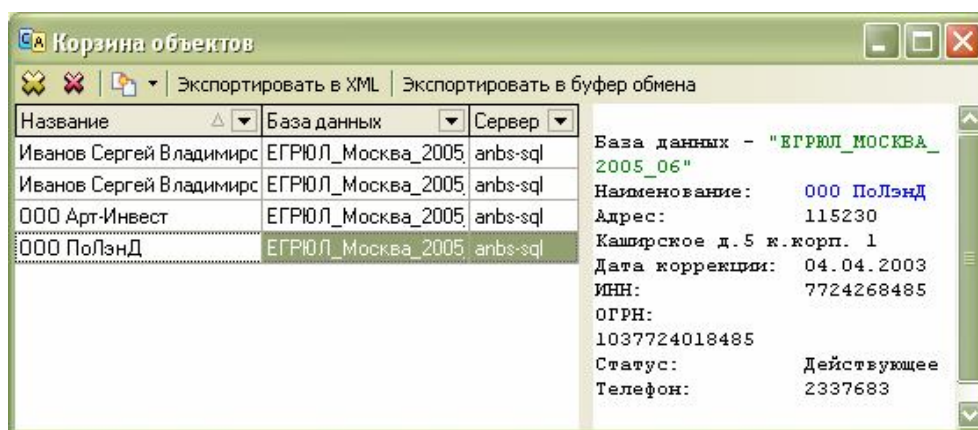
Окно «Мастер поиска»

Найденные данные выделяем и помещаем в корзину объектов при помощи кнопки  «Добавить экземпляр(ы) в корзину», выбирая из списка, в соответствии с результатом поиска, один из способов добавления.



Добавление результатов поиска

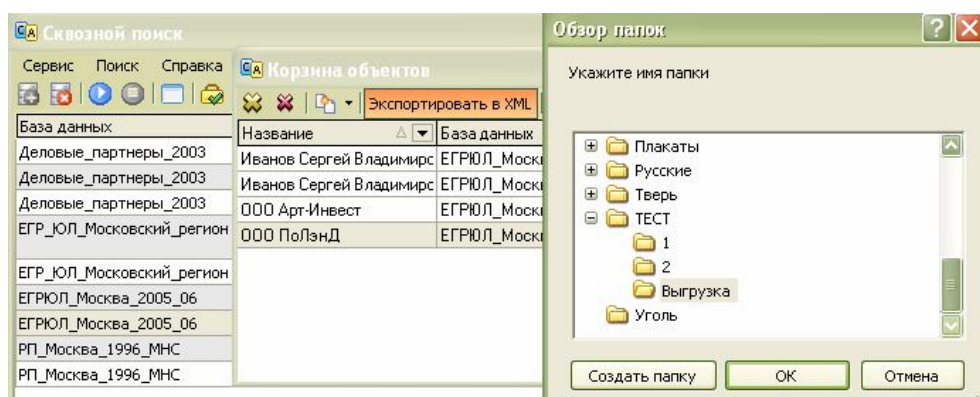
Добавленные экземпляры отобразятся в корзине объектов, которая открывается нажатием на соответствующую кнопку  «Корзина» в левом нижнем углу экрана.



Окно «Корзина объектов»

Для перемещения найденных экземпляров вместе со связанными объектами в АРМ «Аналитик» воспользуемся экспортом в XML файл, который осуществляется путем нажатия соответствующей клавиши в корзине объектов. Сохраняем файл по выбранному вами адресу для дальнейшей его вставки в АРМ «Аналитик». Файл будет иметь наименование, соответственное наименованию БД, например, anbs-sql.ЕГРЮЛ\_Москва\_2005\_06.xml.




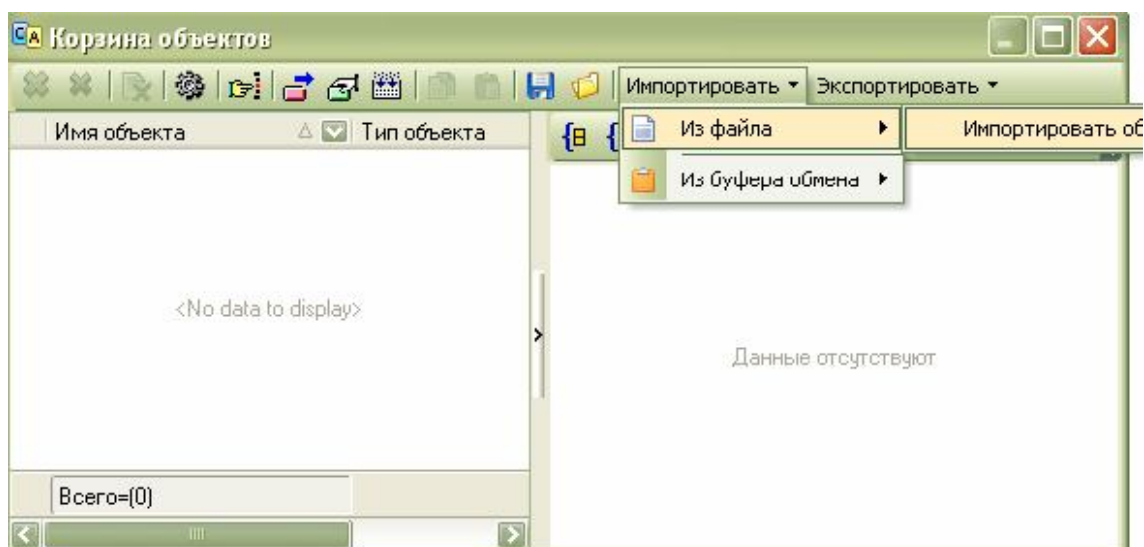


Экспортирование в XML

#### 4.4 Приём данных в АРМ «Аналитик»

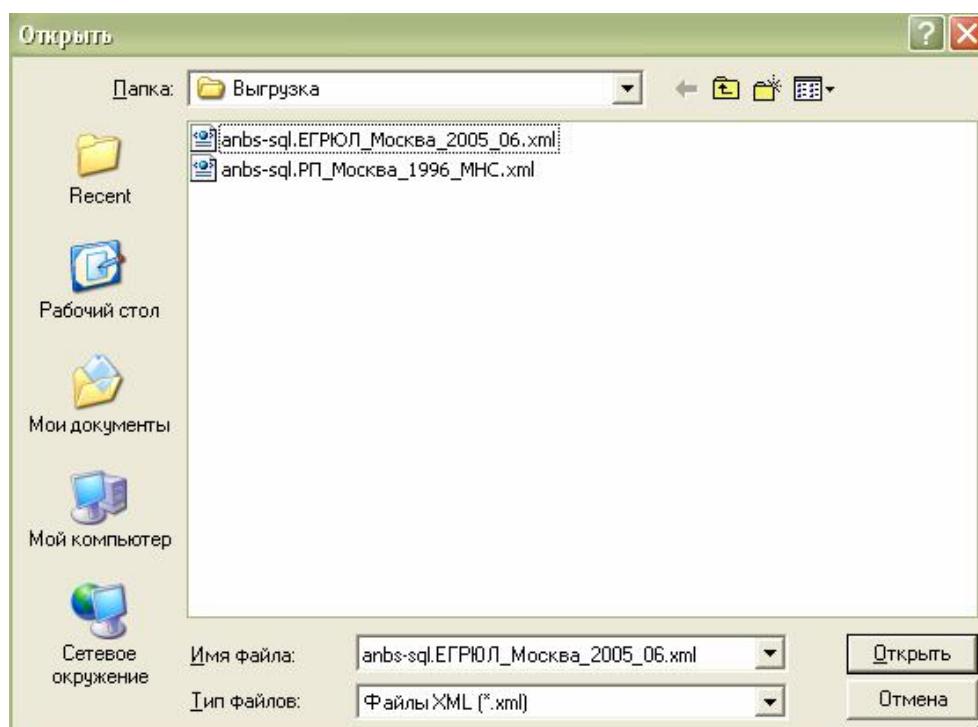
Переходим на витрину АРМ «Аналитик». Для этого последовательно выберите левой кнопкой мыши пункты меню «Пуск» - «Все программы» - «Аналитические бизнес решения» - «Семантический Архив» - «АРМ «Аналитик».

Открываем корзину объектов кнопкой  «Корзина» в левом нижнем углу экрана. Последовательно нажимаем кнопки, как показано на рисунке.



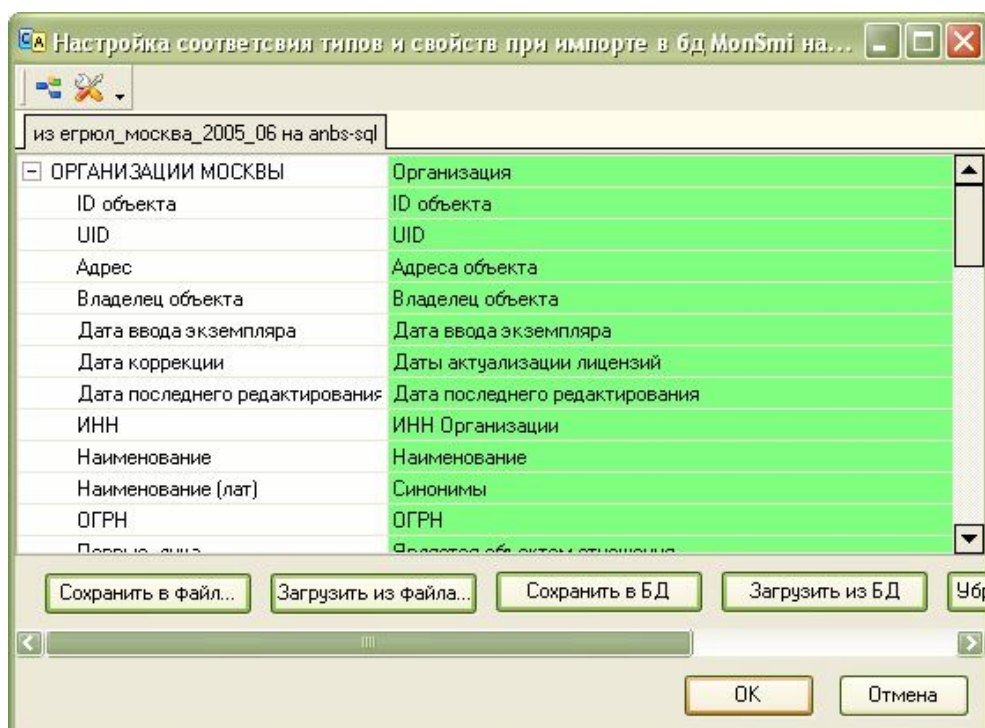
Импорт данных в АРМ «Аналитик»

В открывшемся диалоговом окне зайдите в папку, где хранится XML файл. Нажмите кнопку «Открыть».



Окно «Открыть»

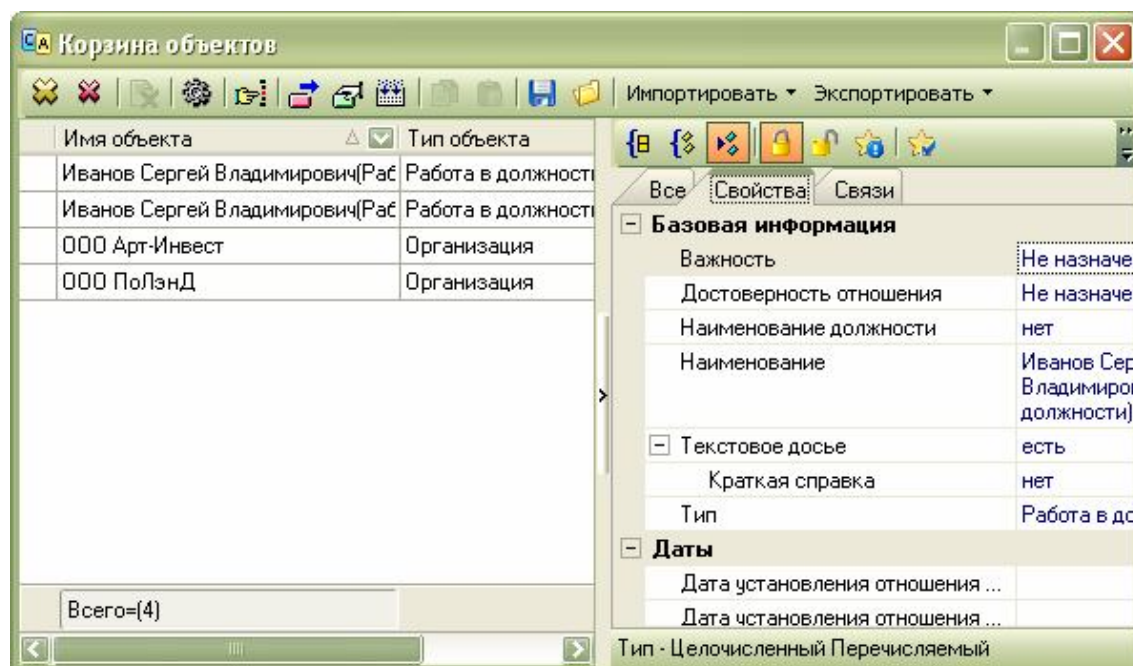
Откроется окно настройки соответствия типов, где по умолчанию предложены соответствия свойств. При необходимости их можно корректировать.



Настройка соответствий типов и свойств при импорте

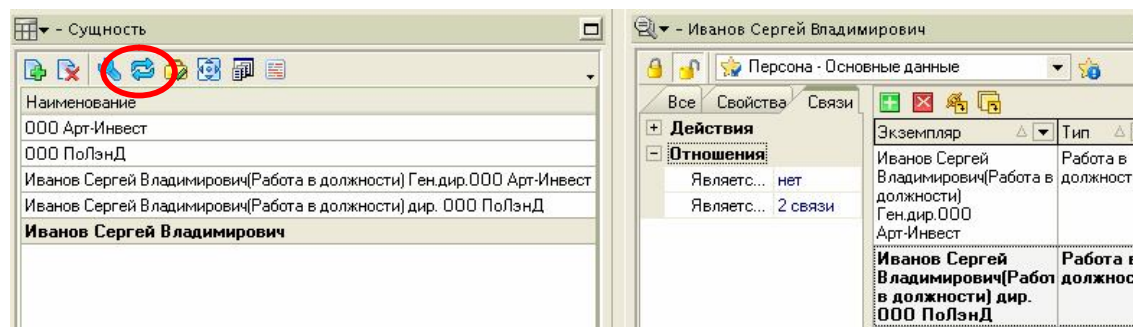
Нажмите кнопку «ОК».

В результате вы увидите корзину объектов с импортированными данными.



Корзина объектов с импортированными данными

Объекты помещены в корзину АРМ «Аналитик», т.е. добавлены в БД Primer\_Persona. Закройте корзину и обновите данные нажатием на кнопку «Обновить». Данные из БД будут привязаны к персоне через отношение «Работа в должности» (т.к. в БД указаны руководители организаций). Вид витрины после выполнения операции показан на рисунке.



Витрина АРМ «Аналитик» после добавления данных из «Корзины»

### *4.3. Задание*

1. Найдите объект «Иванов Сергей Владимирович» через «Сквозной поиск».
2. Переместите данные о нём в АРМ «Аналитик».

### *Контрольные вопросы*

1. Какое средство реализовано в «Семантическом архиве» для поиска данных об объекте?
2. Какое средство реализовано в «Семантическом архиве» для передачи данных об объекте?
3. Файлы какого типа используются для экспорта данных?

## Лабораторная работа №5. Построение семантических сетей

### 5.1. Основная цель


Научиться строить семантические сети с помощью АРМ «Аналитик»".


### 5.2. Пояснения к выполнению работы

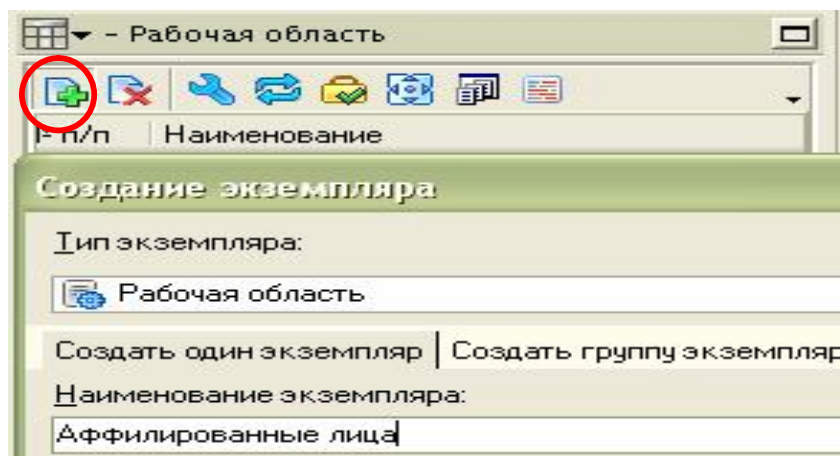
Визуальное отображение данных исследования позволяет наиболее ясно увидеть картину. Проследить прямые, косвенные связи, выявить причинно-следственные взаимодействия. Такая картина сама должна быть максимально ясной и логически интуитивно понятной. Временные процессы следует располагать слева направо, иерархии строить сверху вниз. Процессы, являющиеся подпроцессами одного единого, – объединять рамками. Следует избегать множественного пересечений связей. Связь один ко многим следует отображать одной связью, объединив объекты рамкой.

Анализ собранных данных по объекту начинается с изучения связанных с ним фактов в инспекторе свойств или на динамической раскладке семантической сети. Удобно разбираться в данных, создав экземпляр рабочей области и формируя раскладки с фрагментами данных. Далее в такие раскладки постепенно добавлять новые «куски» сети.

В витрине АРМ «Аналитик» создайте «Рабочую область» «Аффилированные лица».


Для этого нажмите на панели инструментов кнопку  «Поиск элементов по типу».


При помощи кнопки  «Создание экземпляра» создайте экземпляр типа «Рабочая область» с заданным наименованием. Нажмите кнопку «ОК».

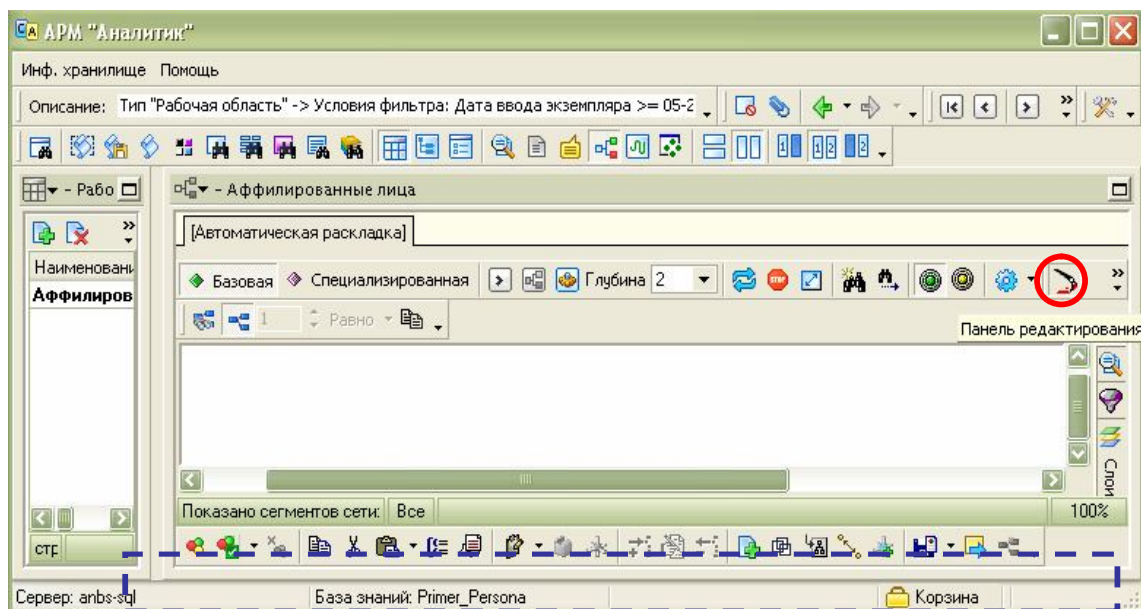


Создание экземпляра в «Рабочей области»

Для перехода в режим «Семантическая сеть» выберите данный тип визуализатора в витрине «Аналитика».

Создана новая рабочая область для построения семантической сети. Перед началом работы нажмите кнопку  «Выделение элементов», т.к. по умолчанию стоит настройка защиты от редактирования.



Для вызова панели редактирования нажмите кнопку . В нижней части экрана появится панель редактирования, которая необходима для построения семантической сети.

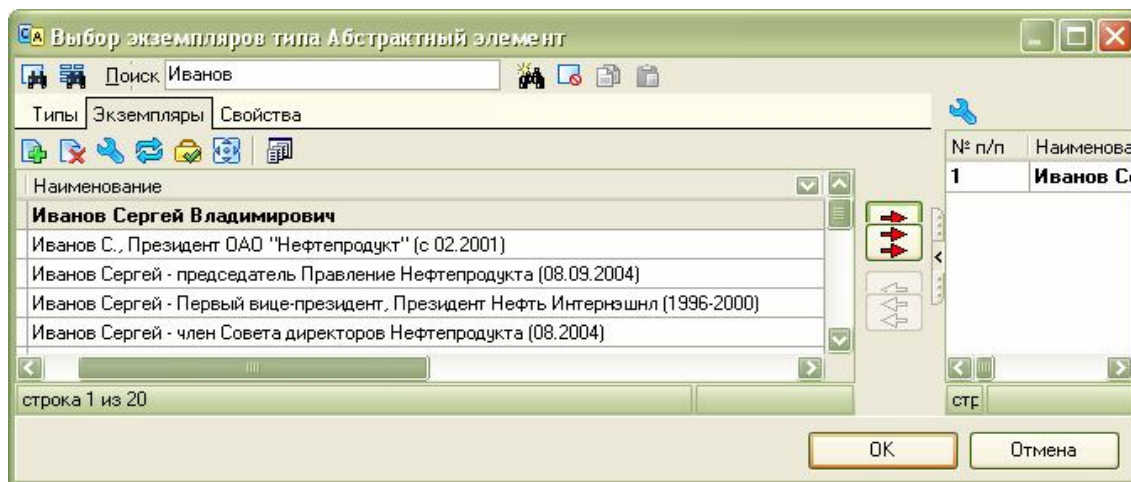


Панель редактирования для построения семантической сети

Начинаем строить семантическую сеть с добавления главного объекта, в нашем примере «Иванов Сергей Владимирович». Нажмите

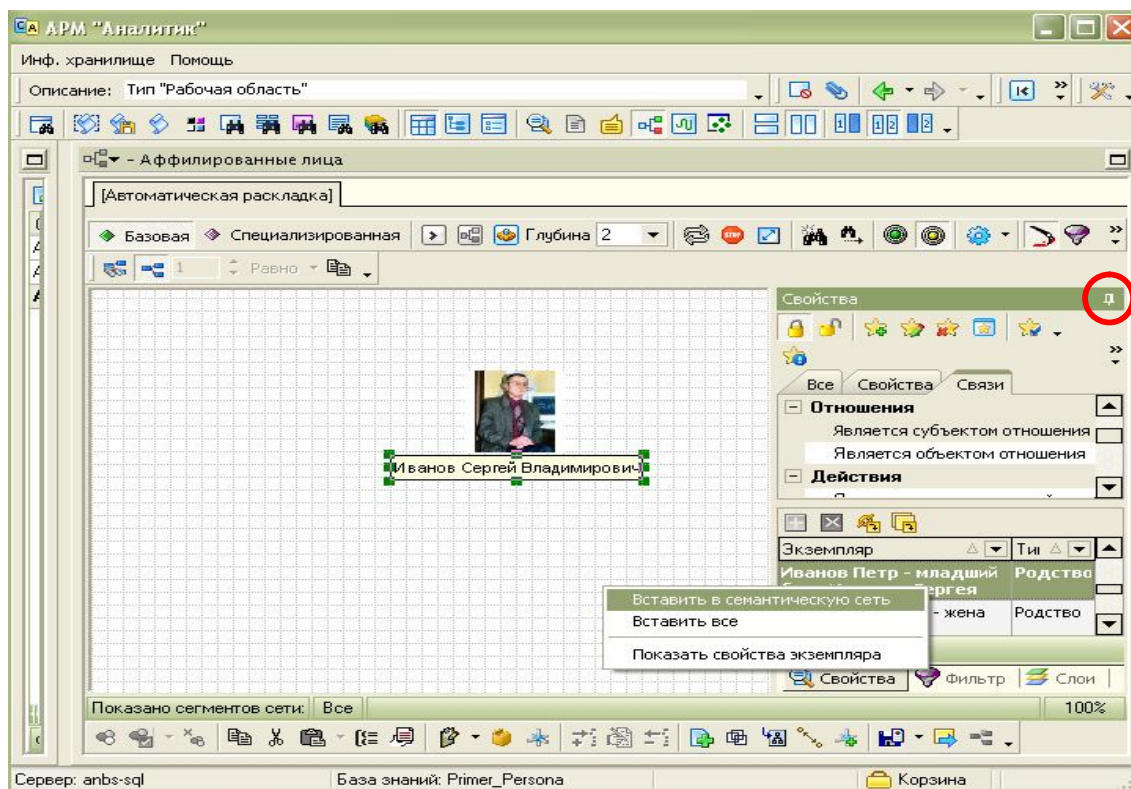


кнопку  «Добавить объект» на панели редактирования. Откроется стандартная панель выбора экземпляров. В строке поиска наберите «Иванов» и нажмите кнопку  «Поиск по имени». Из найденных экземпляров выберите нужный, с помощью стрелочки переместите его в правую часть окна и нажмите кнопку «Ок».



Выбор экземпляра «Иванов Сергей Владимирович»


Экземпляр «Иванов Сергей Владимирович» отобразится на семантической сети. Далее выделите объект (щелчком мыши), справа находится всплывающая панель инспектора свойств. Чтобы закрепить ее, щелкните по скрепке.




Закрепление панели инспектора свойств

На закладке «Связи» отображены все связанные с объектом экземпляры. Выделите один или несколько экземпляров и нажатием правой клавишей мыши выберите «Вставить в семантическую сеть».

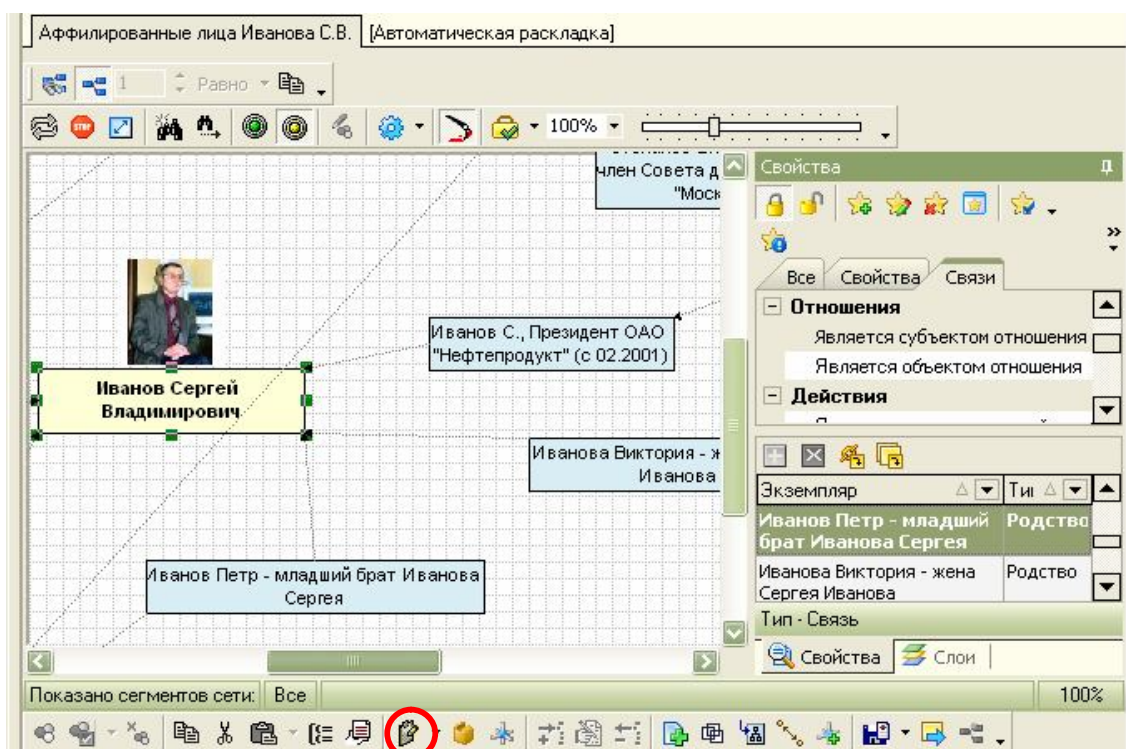
Таким образом, постепенно наполняется семантическая сеть, распределяются экземпляры.

Для сохранения результата работы воспользуйтесь кнопкой  «Сохранить».

Дайте наименование раскладке, например «Аффилированные лица Иванова С.В.».

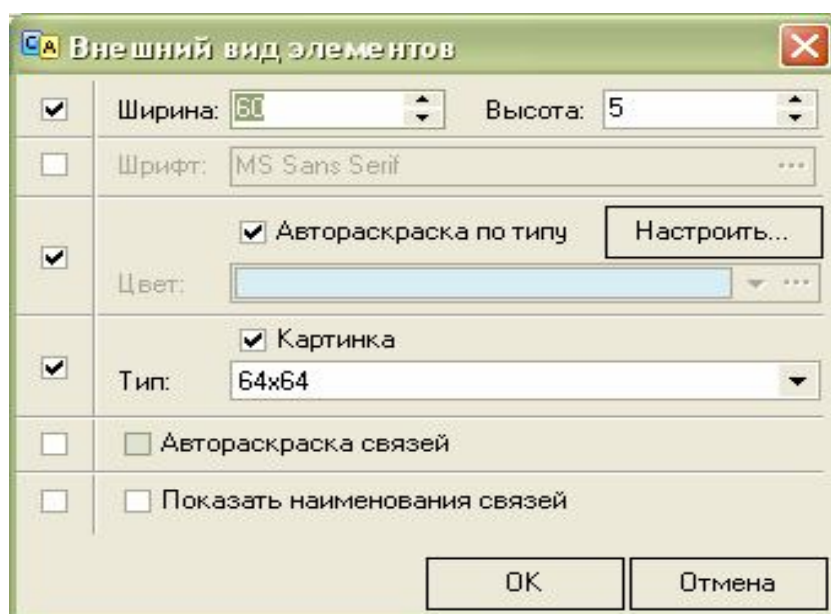
Для придания раскладке эстетичного вида воспользуйтесь кнопкой  «Форматирование».





Открытие панели «Форматирование»

Из выпадающего списка выберите пункт «Внешний вид элементов». Откроется окно «Внешний вид элементов».

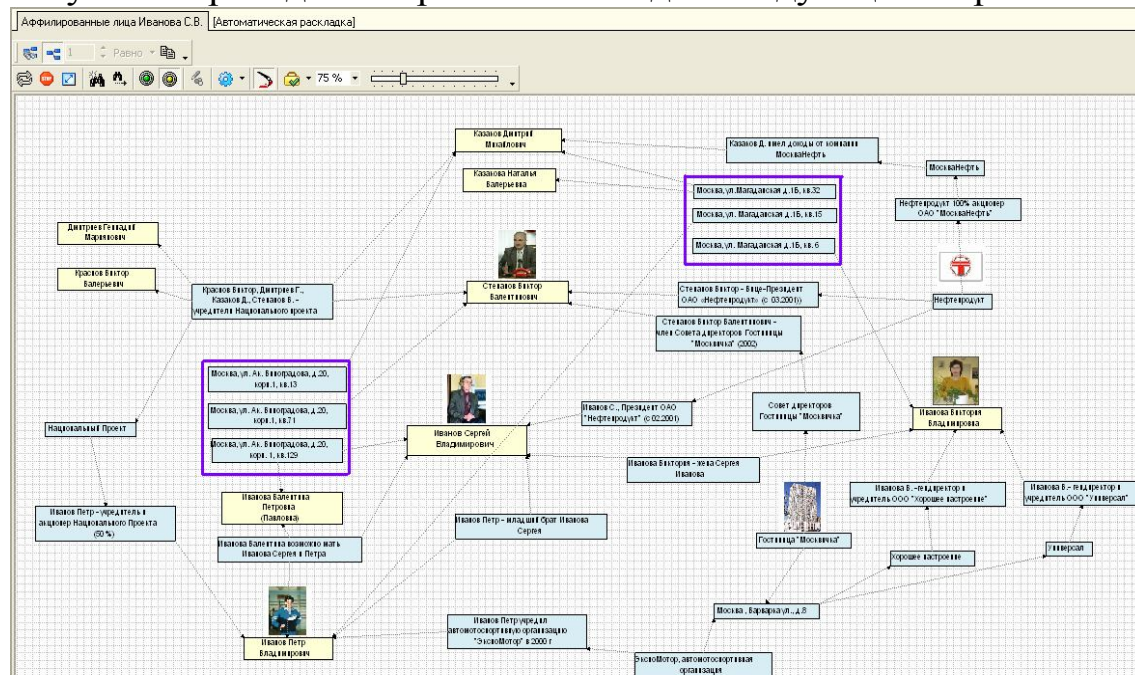


Окно для настройки внешнего вида элементов


В этом окне можно редактировать внешний вид экземпляров: изменять размеры, цвет и т.д.

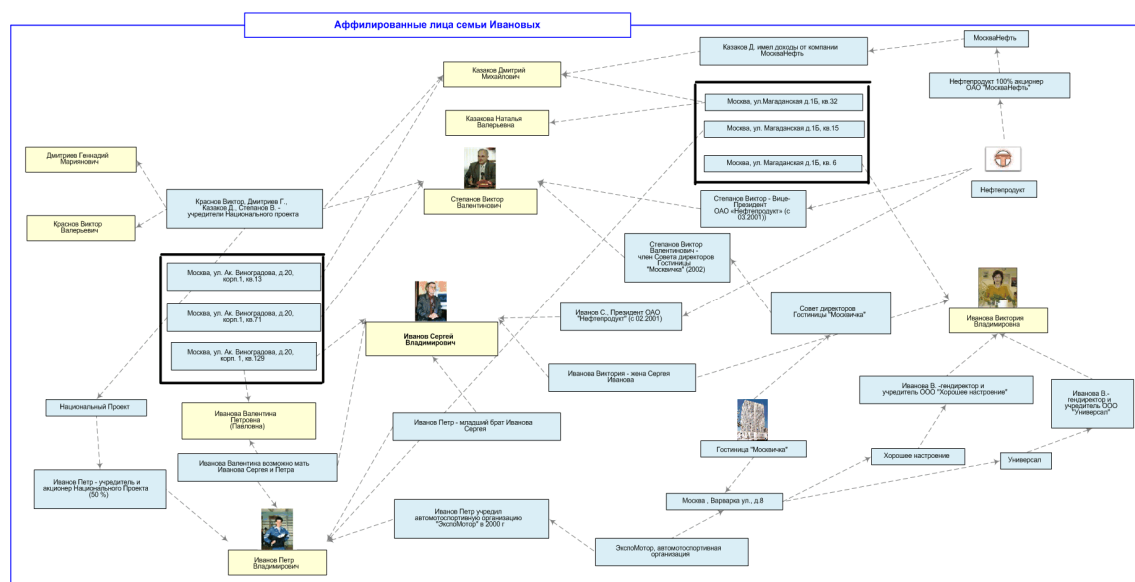
По умолчанию настроены цвета в соответствии с типом объекта.

Результат проведенной работы выглядит следующим образом.



Семантическая сеть для объекта «Иванов С.В.»

Чтобы сохранить раскладку в формате программы MS Visio, нажмите кнопку  «Экспорт в VISIO». Последовательно выполняйте шаги мастера экспорта. Вы получите изображение в формате «.vsd».



Семантическая сеть в формате «.vsd»

Можно воспользоваться шаблоном оформления заголовка и легенды, либо создать свой фирменный шаблон и экспортировать раскладки в него.

На семантической сети есть возможность расположения экземпляров в виде таблицы.

### *5.3. Задание*

1. Запустите АРМ «Аналитик».
2. Найдите в БД объект для построения семантической сети.
3. Создайте рабочую область «Аффилированных объектов».
4. Постройте семантическую сеть.
5. Экспортируйте полученный результат в MS Visio.

### *Контрольные вопросы*

1. Какие типы объектов могут быть использованы для построения семантической сети?
2. Как можно сохранить полученную сеть?
3. Какие существуют способы расположения экземпляров на семантической сети?

## **Лабораторная работа №6. Формирование дайджестов статей**

### *6.1. Основная цель*

Научиться формировать дайджесты статей в АРМ "Аналитик".

### *6.2. Пояснения к выполнению работы*

Визуальное отображение картины исследования не даёт возможности полного представления данных за счёт отображения только ключевых слов. Чтобы представлять полную, целостную картину, необходимы текстовые расшифровки в виде кратких справок из первоначальных источников, которыми служат статьи и заметки СМИ.

### *6.3. Настройка содержания дайджеста статей*

Дайджест (от англ. digest) – краткое изложение, резюме.

Разделение статей на тематические блоки:

- если в свойствах статей «Тематический блок» не указан, все статьи будут относиться в содержании к «Разное»;
- если требуется четкое разделение статей по тематике, тогда необходимо связать статьи, связанные тематикой.

Статьи, которые необходимо объединить тематическим блоком, выделяются и помещаются в «Корзину».

№ п/п	Наименование	Добавить в корзину
50	Партнером сети кофеен Starbucks станет компания "Интерспорт Россия"	
51	«АвтоВАЗ» не смог договориться с General Motors о выпуске Opel Corsa	
52	<b>Chery Automobile и "Автотор" подписали соглашение</b>	
53	На "Автоторе" началось производство Hummer H3	
54	07.06.2006 Выпущено новое авто под названием "КингКонг"	
55	Глава Роспрома Борис АЛЕШИН: Корейский автомобиль - это чуть ли не высшая оценка	
56	"Автотор" получил по "отвертке"	
57	"Капсулы смерти" реются в Россию: Китайские автомобили продаются у нас все лучше	
58	"Автотор" может лишиться льгот по всем контрактам	
59	Chery сошел с конвейера в России	
60	"Богдан" подкатит в Россию еще Chevrolet. Украинцы хотят построить автозавод под Ник	
61	ЗАЗ и "Богдан" получили вид на автомобилестроительство. Завершилась сделка по покуп	
62	Богдан" и ЗАЗ наезжают на Россию	
63	Opel и Chrysler будут выпускать в России	
64	Chery построит завод в России без участия "Автотора"	
65	«АвтоВАЗ» растет, иномарки в дефиците	
66	"Автотор" подошел к Kia сее'd SW "универсально"	
67	Lada поехала	
68	Калининград - под цунами автовойн	
69	«Сид» - двадцать второй	
70	Международный Московский Банк – официальный спонсор «Интеравто-2007»	
71	За счет иномарок	
72	КТО КРЕПЧЕ ДЕРЖИТ РУЛЬ	
73	Опасный импорт	
74	150 калининградских контейнеровозов "играют в прятки с ГИБДД" из-за волокиты чинов	
75	Больше китайцев	

Добавление статей к тематическим блокам

Затем, открыв «Корзину», нажимается кнопка «Работать как с одним».

Корзина объектов	
Имя объекта	Тип объекта
"Капсулы смерти" реются в Р	Статья
07.06.2006 Выпущено новое а	Статья
Chery Automobile и "Автотор" г	Статья
Chery построит завод в Росси	Статья
Chery сошел с конвейера в Ро	Статья

Всего=[5]

Импортировать Экспортировать

Все Работать как с одним

Все связи 8 связей

**Служебные**

Банк данных	есть
Знания выделены	Да
Оператор	Г.Бекмуратов
Статус автоматической...	3
Дата ввода экземпляра	14.09.2007
Источник содержит Эл...	8 связей

**Реквизиты статьи**

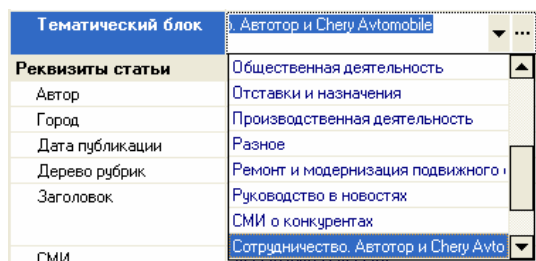
Заголовок	"Капсулы смер Китайские авто все лучше
Дата публикации	13.08.2007 13:4
СМИ	Steer.ru (http://s
Автор	есть
Город	Не назначен
Язык статьи	Не назначен

Тип - Связь

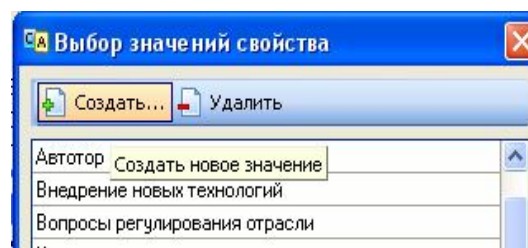
Кнопка «Работать как с одним»



Далее указывается нужный тематический блок или создается новый.



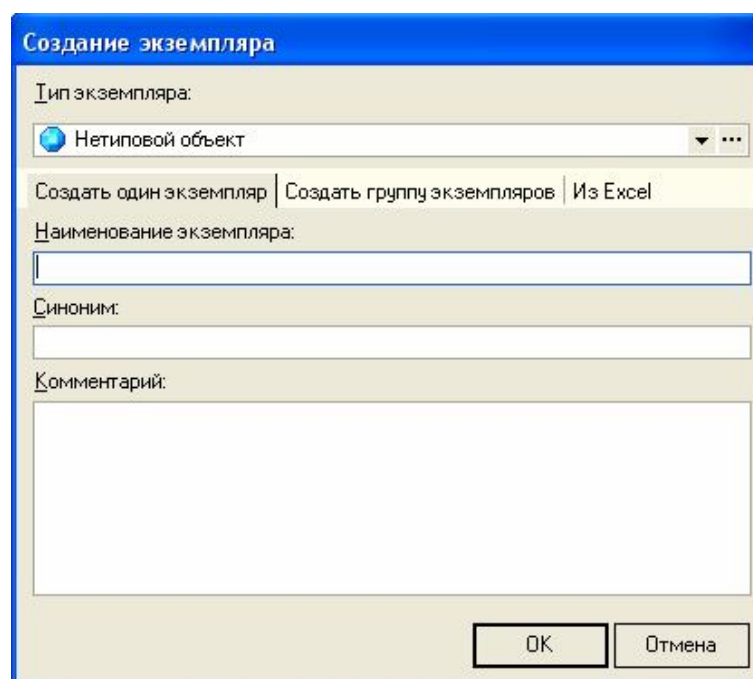
Указание тематического блока



Создание нового тематического блока

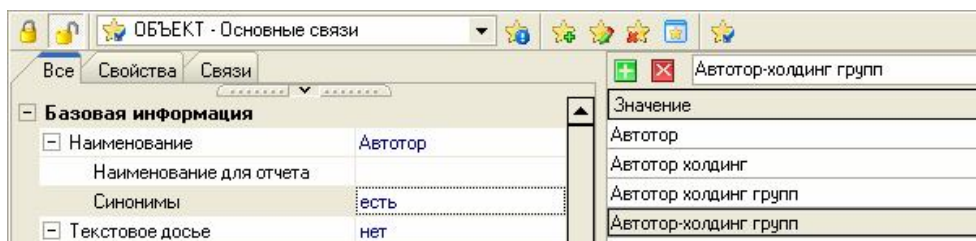
### *Подсветка ключевых слов и подсчет статистики по этим словам*

Для того чтобы в дайджесте статей автоматически выделить ключевые слова, необходимо создать объекты для этих ключевых слов (желательно для каждого нового слова отдельный объект).



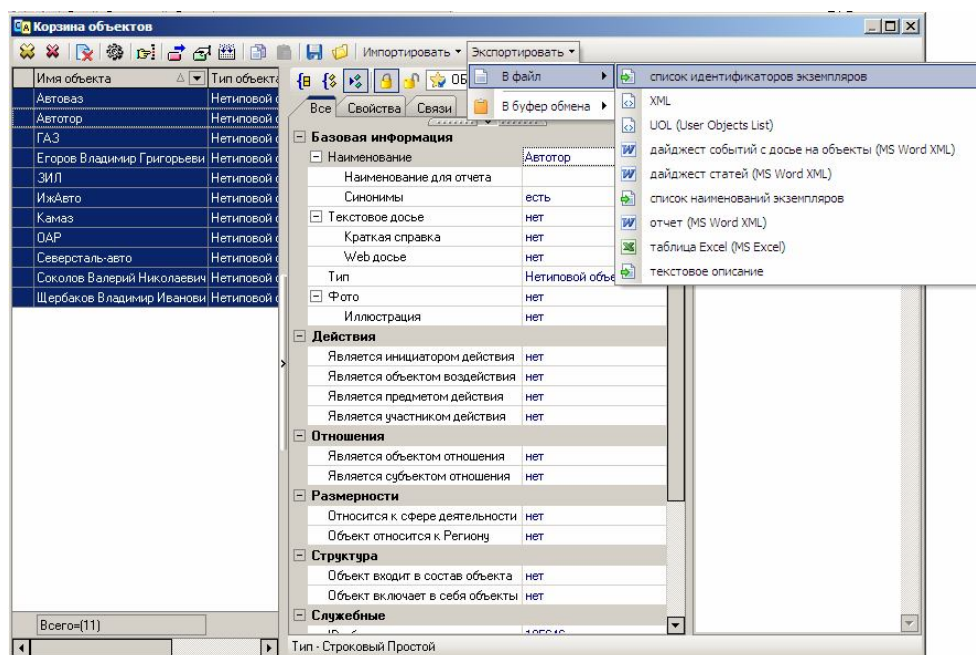
Создание объектов для ключевых слов

Затем в свойствах созданного объекта необходимо прописать синонимы, по которым этот объект может встречаться в тексте.



Добавление синонимов

После объект или группа объектов помещается в «Корзину» и экспортируется в список идентификаторов экземпляров.

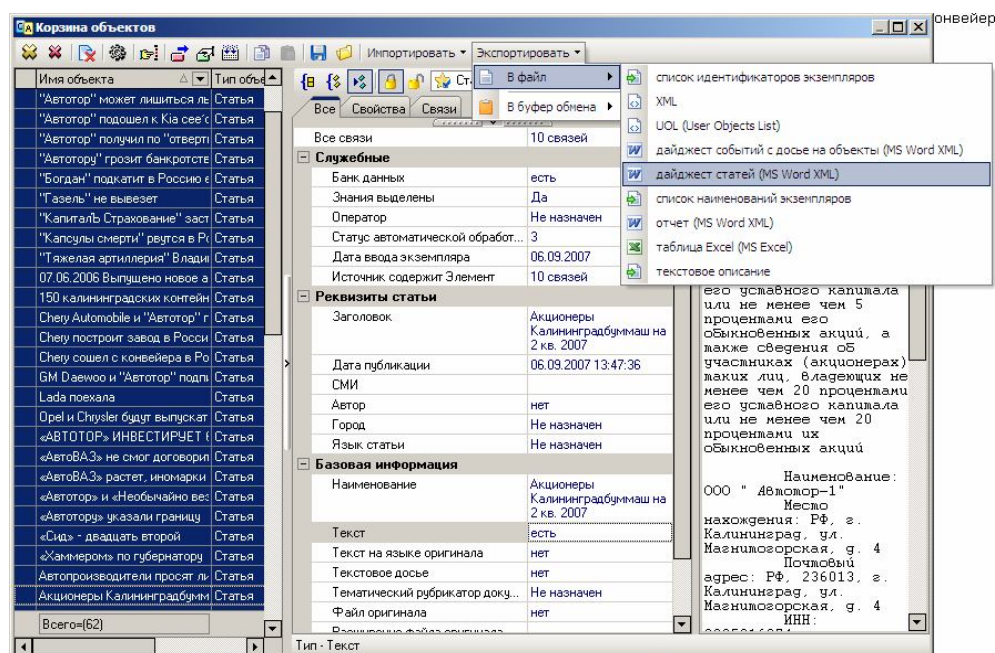


Экспорт объектов в список идентификаторов экземпляров

Список идентификаторов экземпляров сохраняется по умолчанию в папку C:\Program Files\ABS\Semantic Archive\Reports, но этот путь можно изменить.

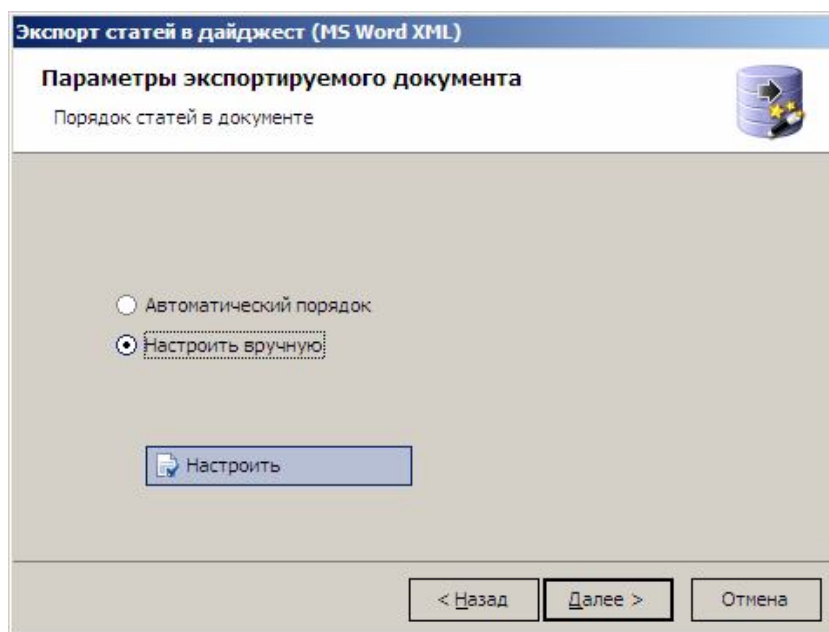
### Экспорт статей в дайджест

После выделенные для дайджеста статьи помещаются в «Корзину» и экспортируются в дайджест статей.



Экспорт статей в «Дайджест статей»

Далее появится окно «Экспорт статей в дайджест». Вводим параметры экспортируемого документа и настраиваем содержание.

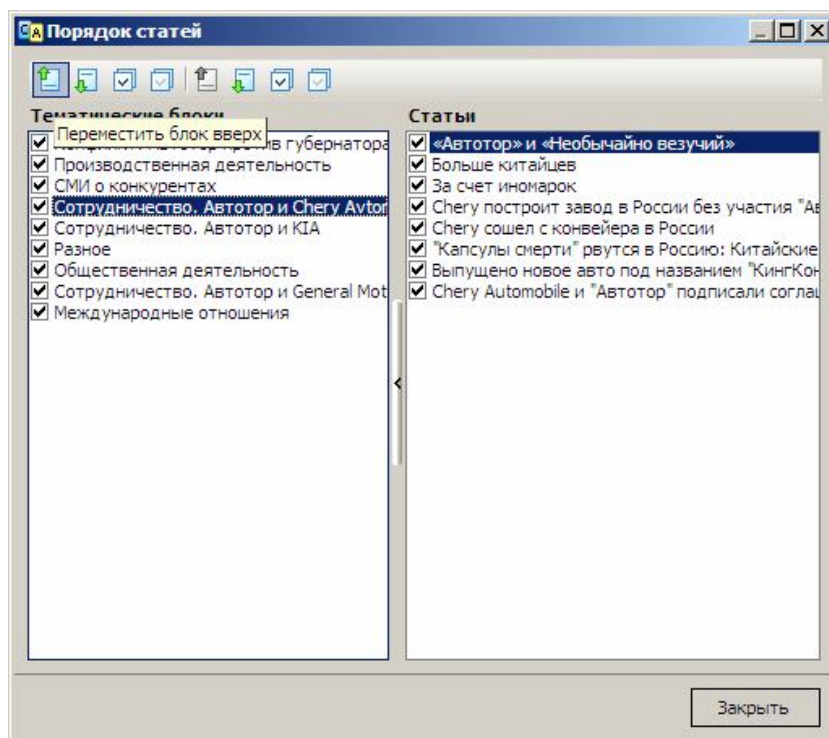


Настройка содержания дайджеста



## Настройка Экспорта статей в дайджест

Далее появится окно «Порядок статей». При желании можно передвинуть заголовки содержания по степени важности.

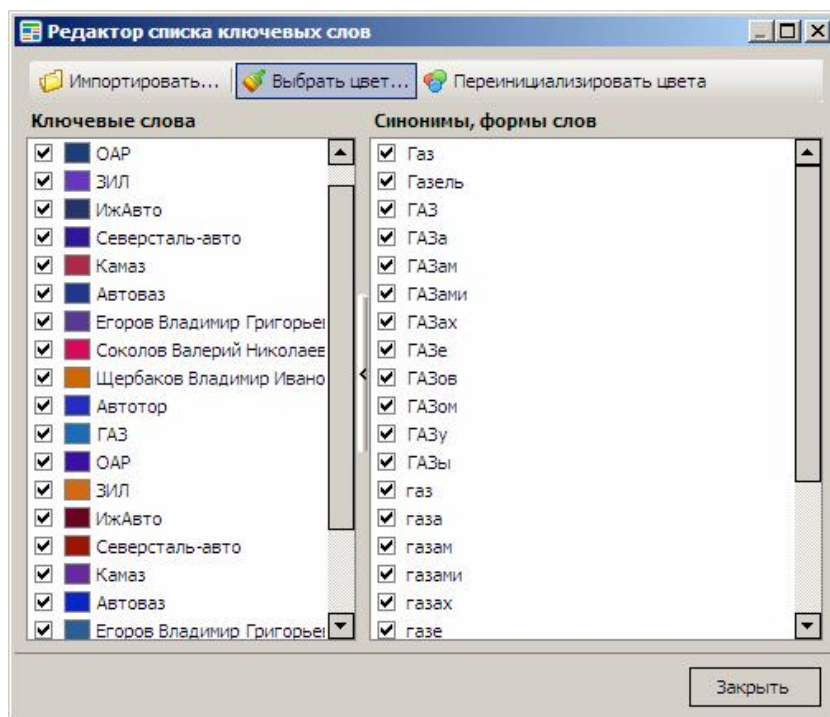


Перемещение заголовков

Затем настраивается подсветка и подсчет (статистика упоминаний ключевых слов).

Ключевые слова импортируются из папки C:\Program Files\ABS\Semantic Archive\Reports.

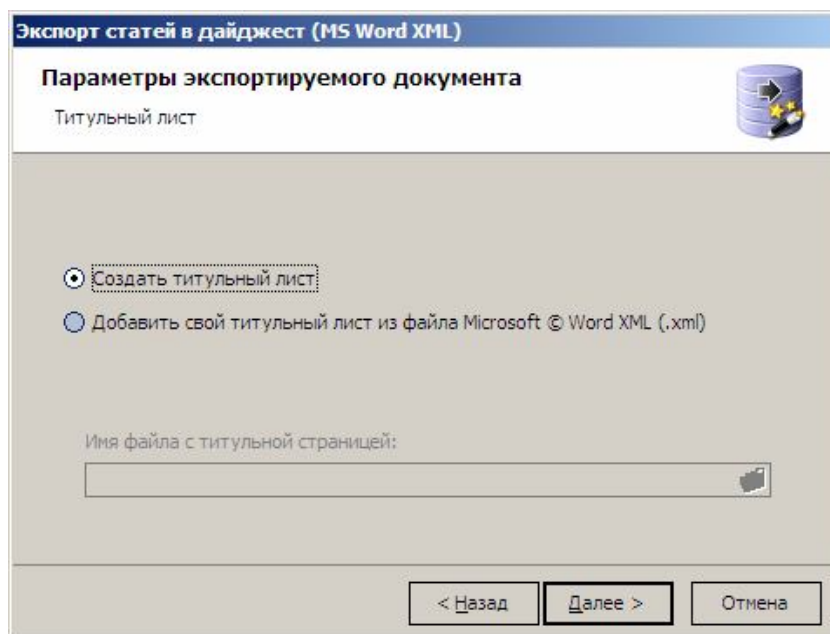
Нажав на кнопку «Список ключевых слов», можно изменить их параметры.



Редактор списка ключевых слов

Далее выбирается нужный шаблон дайджеста. При желании так же можно указать параметры документа (размер и тип шрифта, стиль и пр.).

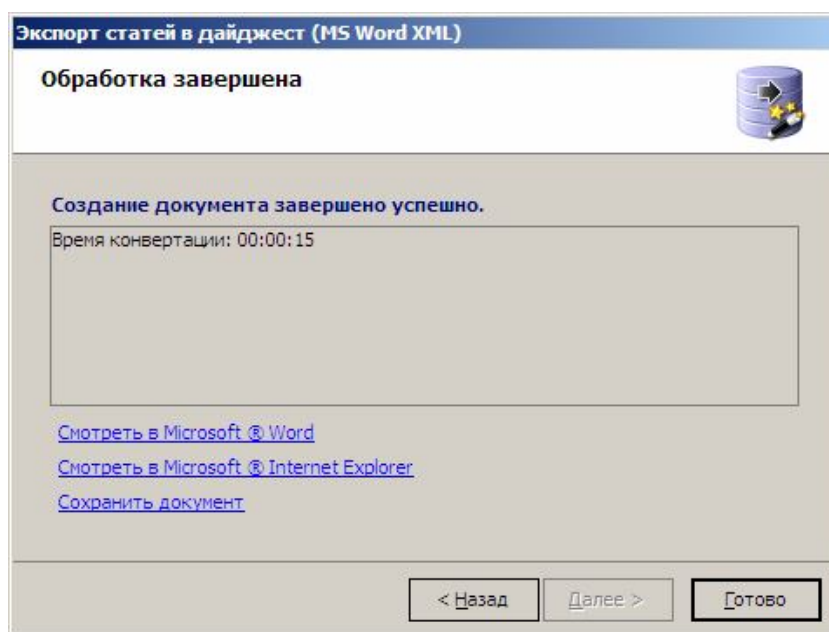
Если есть титульный лист к дайджесту, можно его указать системе, чтобы она автоматически его вставила.



Настройка титульного листа

Если в дальнейшем вам будут нужны эти же настройки дайджеста, можно их сохранить и в дальнейшем вам не придется снова их указывать, а просто указать в начале генерации отчета шаблон дайджеста.

После того как дайджест сгенерирован, над ним можно будет или просмотреть в MS Word, или в Internet Explorer, или сразу сохранить его.



Окно завершения обработки

Дайджест статей готов.

ФИО Иванов Сергей Владимирович  
Дата рождения 15.09.1961  
Пол Муж  
Наименование типа Персона  
Национальность Русский



#### Текстовое досье

Сергей Владимирович Иванов родился 15 сентября 1961 г. В 1983 г. окончил Сургутский нефтяной техникум; в 1990 г. – Гроенский институт нефти и химии по специальности "Бурение нефтяных и газовых скважин", квалификация - инженер. В 1979-1981 гг. – проходил срочную военную службу. После окончания техникума работал на обустройстве месторождений Ямала специалистом треста «Ямалнефтегазсервис», мастером участка, мастером специализированного участка треста «Ямалнефтедормонтаж». С 1990 года возглавлял компанию «Иванов и партнеры» (Москва), с 1993 года - "ОйлтрансТрейддинг" (исполнительный директор). С 1996 года - на руководящих должностях в нефтяной компании «Ойлтранс». С ноября 2000 г. - Президент ОАО "Нефтяной консорциум".

#### Упоминания в СМИ

Сергей Иванов возглавил "Нефтяной консорциум" в конце 2000 года по рекомендации Петра Трофимова. «Он предложил мою кандидатуру, сказал, что рукою за меня как за трудоспособного человека и специалиста», - вспоминает Иванов в интервью Forbes. Вместе с Трофимовым он несколько лет проработал в компании «Ойлтранс», а на важный пост крупнейшей российской нефтяной компании попал благодаря ее бесновому руководителю Николаю Гаеву. [СМИ: Ведомости, Дата публикации: 01.08.2004, Заголовок: Назначения]

У певицы Светланы Ионовой отношения с солидным мужчиной Сергеем Ивановым достаточно серьезные. Президент нефтяной компании заливает ее дорогими подарками. Первый же презентом стало платье стоимостью 10000 долларов от Versace, а потом Иванов подарил певице квартиру на северо-западе столицы. [СМИ: Московский комсомолец, Дата публикации: 05.07.2004, Заголовок: Певица сменила мужчину]

#### Родственные связи

Жена - Иванова Анна Владимировна (11.12.1978 г.р.) (девичья фамилия Копылова)  
Дочь - Иванова Мина Сергеевна  
Сын - Иванов Федор Сергеевич  
Женат вторым браком.  
/В первой семье - тоже двое детей, жена Иванова Наталья Михайловна (1963 г.р.) - непроверенные данные/  
Брат - Иванов Петр Владимирович (10.07.1966)  
Мать (возможно) - Иванова Валентина Петровна (Павловна) (03.10.1939)  
[В 2002 г. доход Ивановой В.П. от Благотворительного фонда "Ойлтранс" составил 1988,706 тыс. руб.]

#### Увлечения, хобби

Сергей Владимирович увлекается восточными единоборствами.

## Нефтяной консорциум

Тип: Коммерческая организация  
Дата основания не позднее: 04.08.1994



Адрес филиалов/офисов	Россия, г. Москва, ул. Садовая-Самотечная, д. 121, строение 6
Тел.	(495) 277-38-80, 277-74-71
Факс	(495) 277-38-80
Руководитель	Иванов Сергей Владимирович, Президент
Деятельность	Поиск месторождений, геологическое изучение, разведка, добыча углеводородов
Объем продаж	8066 711 000 руб. (за 01.01.2001)
Нераспределенная прибыль	0 руб.
Численность персонала	7540 человек
Активы	3 989 298 000 руб. (за 01.01.2001)

#### Краткая справка

Полное фирменное наименование: Открытое акционерное общество «Нефтяной консорциум производителей нефтепродуктов».  
Сокращенное наименование: ОАО "Нефтяной консорциум".  
Место нахождения: Россия, г. Москва, ул. Садовая-Самотечная, д. 121, строение 6.  
Почтовый адрес: Россия, г. Москва, ул. Садовая-Самотечная, д. 121, строение 6.  
Доля эмитента в уставном капитале хозяйственного общества: 100 %.  
Данное хозяйственное общество является по отношению к эмитенту дочерним.  
Доля данного лица в уставном капитале эмитента: доли не имеет.  
Основной вид деятельности: поиск месторождений и производство нефтепродуктов.

#### Упоминания в СМИ

"Нефтяной консорциум" открыл представительство "Нефтяной консорциум" (ОАО) в г. Нижнеартовск. ОАО "Нефтяной консорциум" рассматривает г. Нижнеартовск в качестве региона, имеющего значительный экономический и промышленный потенциал. Главная задача Представительства – установление и развитие эффективного сотрудничества с предприятиями и учреждениями Нижнеартовска.

[СМИ: "Нефтяной консорциум" (ОАО), Дата публикации: 23.11.2005]

На днях руководство компании Нефтяной консорциум выступило с общим обзором достижений за прошедший год. Результаты впечатляют: компания имеет один из лучших показателей по нефтедобыче, стала крупнейшим инвестором в долгосрочные перспективные проекты, активно развивает производственные подразделения именно в России. Только рост добычи нефти и газа на предприятиях, входящих в компанию, за год вырос на 14%. [СМИ: Российская газета, Дата публикации: 07.09.2003, Заголовок: "Нефтяной консорциум на подъеме"]

#### Акционеры

На конец 1999 года  
Минимому количеству передано государством 25,5 % простых акций ОАО «Нефтяной консорциум» (120% - для передачи на баланс ПО "ОКОНА")

## Дайджест статей

Для распечатки дайджеста следует нажать кнопку  «Печать».

### 6.3. Задание

1. Запустите APM "Аналитик".
2. Найдите в базе данных статьи на любую тематику.
3. Разделите статьи на тематические блоки.
4. Выделите интересующие объекты.
5. Сформируйте дайджест статей.

### Контрольные вопросы

1. Для чего применяется объект «Корзина»?
2. Для чего используются ключевые слова?
3. В какие форматы можно экспортировать дайджест статей?
4. С помощью каких программ можно просмотреть сохранённый дайджест?

### Библиографический список

1. Методы и модели анализа данных: OLAP и Data Mining. / *А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод* – СПб.: БХВ-Петербург, 2004.- 336 с.: ил.
2. *А.В. Бурьяк*. Аналитическая разведка [Электрон. ресурс] / *А.В. Бурьяк*. - 2008.- Режим доступа: <http://analytical.narod.ru/>
3. *Джексон П.* Введение в экспертные системы : [пер. с англ.] / *П. Джексон*. - 3-е изд. - М.: издательский дом «Вильямс», 2001. – 624 с.
4. Аналитика: методология, технология и организация информационно - аналитической работы /*П.Ю. Конотопов, Ю.В. Курносков* - М.: РУСАКИ, 2004. - 512 с.
5. *И. Ю. Нежданов*. «Аналитическая разведка для бизнеса». - М.: издательство «Ось-89», 2008.
6. Официальный сайт компании «BaseGroup Labs» [Электрон. ресурс]. Рязань, 1995-2010.- Режим доступа: <http://www.basegroup.ru/>
7. Официальный сайт компании «Аналитические бизнес решения» [Электрон. ресурс]. М.- Режим доступа: <http://www.anbr.ru/index.php?lang=1>
8. Data Mining и аналитическая платформа Deductor [Электрон. ресурс] : [статья]. М., 2008.- Режим доступа: [http://stt-s.ru/index.php?option=com\\_content&task=view&id=56&Itemid=90](http://stt-s.ru/index.php?option=com_content&task=view&id=56&Itemid=90)
9. *Г.И. Просветов* (МГУ им. М.В. Ломоносова). Дерево решений [Электрон. ресурс] : [статья] / *Г.И. Просветов*.- СПб, 2008.- Режим доступа: [http://www.elitarium.ru/2008/04/09/derevo\\_reshenijj.html](http://www.elitarium.ru/2008/04/09/derevo_reshenijj.html)
10. *Дюк В.А.* (СПИИРАН). Data Mining – интеллектуальный анализ данных [Электрон. ресурс] : [статья] / *В.А.Дюк*. – СПб. - Режим доступа: <http://www.olap.ru/basic/dm2.asp>
11. *Деревянко В.А.* Поиск ассоциативных правил при интеллектуальном анализе данных [Электрон. ресурс] : [статья] / *В.А. Деревянко*.- 2009.- Режим доступа: [http://www.rammus.ru/products/arda/article\\_lam\\_translation](http://www.rammus.ru/products/arda/article_lam_translation)