На правах рукописи

РАБИНОВИЧ БОРИС ИЛЬИЧ

ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ КОМПЛЕКСНОЙ ОБРАБОТКИ ИНФОРМАЦИИ В РАМКАХ ЛОГИКО-АНАЛИТИЧЕСКОЙ СИСТЕМЫ НА ОСНОВЕ РАСШИРЕННЫХ СЕМАНТИЧЕСКИХ СЕТЕЙ

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Москва, 2008

Работа выполнена в Институте проблем информатики Российской академии наук

Научный руководитель:	доктор технических наук, профессор Кузнецов Игорь Петрович
Официальные оппоненты:	доктор технических наук, профессор Шемакин Юрий Иванович
	кандидат технических наук, доцент Башлыков Александр Александрович
Ведущая организация:	Московский технический университет связи и информатики

Защита диссертации состоится _____2008 г. в ___ часов на заседании диссертационного Совета Д002.073.01 при Институте проблем информатики РАН по адресу: 119333, Москва, ул. Вавилова, 44, корп. 2.

С диссертацией можно ознакомиться в библиотеке Института проблем информатики Российской академии наук.

Отзывы в одном экземпляре, с заверенной подписью, просим направлять по адресу: 119333, Москва, ул. Вавилова, 44, корп. 2, в диссертационный Совет.

Автореферат разослан « ___» _____ 2008 г.

Ученый секретарь диссертационного совета Д002.073.01
доктор технических наук, профессор

С.Н. Гринченко

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. В настоящее время наблюдается повсеместный лавинообразный рост потоков разнородной информации, состоящей из сложноорганизованных документов, различных отчетов, электронных писем и пр. В связи с этим актуальным является разработка технологий и программных средств комплексной обработки разнородной информации. В криминальной милиции примером разнородной информации могут быть тексты на естественном языке (сводки происшествий, обвинительные заключения, справки по уголовным делам), данные из различных справочников (телефоны, адреса), биллинги телефонных переговоров и др. Информация может храниться в файлах, в базах данных (БД) или извлекаться из сети Интернет. Её обработка должна быть максимально автоматизирована, что зачастую предполагает решение сложных логико-аналитических задач (поиск объектов, анализ их связей и др.). Перспективным является разработка технологий и систем, позволяющих осуществить на единой основе агрегацию, хранение и логико-аналитическую обработку разнородной информации достаточно унифицированными средствами.

Такая система, ориентированная на обработку текстов на естественном языке (ЕЯ), разработана в Институте проблем информатики Российской академии наук в рамках проекта «Аналитик» и связанных с ним проектов «Криминал», «Дискурс» и «Поток». Созданная система «Аналитик» нашла применение в МВД и ГУВД города Москвы.

Ее особенность заключается в использовании семантико-ориентированного лингвистического процессора, позволяющего отобразить тексты на ЕЯ в структуры знаний, которые образуют базу знаний (БЗ). Для представления информации в БЗ используются расширенные семантические сети (РСС). Их отличие от обычных семантических сетей состоит в использовании многоместных фрагментов, связывающих вершины, и кодов фрагментов, которые тоже являются вершинами. Такие сети позволяют с достаточной точностью представлять объекты и их связи, которые выражаются в ЕЯ с помощью различных форм, в том числе с отглагольными существительными, с оборотами с инфинитивами, со сложноподчиненными предложениями. Связанными могут быть не только объекты, но и сами действия, в которых эти объекты принимают участие.

Обработка информации в системе «Аналитик» осуществляется с помощью языка Декл, созданного для обработки структур знаний в виде РСС. На языке Декл разработано много уникальных программ семантического поиска в БЗ (поиск похожих объектов и ситуаций, поиск по связям и по приметам и др.), программ аналитической обработки и экспертных оценок (семейство оболочек экспертных систем). Использование в качестве БЗ обычных семантических сетей, языков логики предикатов, фреймов приводит к существенной потере информации, содержащейся в текстах на ЕЯ, и, соответственно, к ограничению круга решаемых задач.

Представляется перспективным дальнейшее развитие систем, основанных на структурах знаний в виде РСС. Основными направлениями их развития должны быть: обработка разнородной информации в рамках единой БЗ с использованием уже имеющихся средств; разработка средств решения новых логико-аналитических задач; обеспечение взаимодействия БЗ с внешними БД. В этом случае пользователь-аналитик будет получать из одного источника полную информацию в наиболее удобном виде.

Объекты исследования. Объектами исследования диссертационной работы являются системы, основанные на технологии БЗ, обеспечивающие обработку на единой основе разнородной информации (неструктурированной – текстов на ЕЯ, слабоструктурированной – биллингов, структурированной – данных из БД); существующие методы обработки различных типов данных и их структур; модели информационных структур; методики интеграции данных.

Цель диссертационной работы. Целью диссертации является разработка информационной технологии комплексной обработки разнородной информации большого объема в рамках логико-аналитической системы, основанной на структурах знаний в виде РСС (далее Система).

Основные задачи исследования:

1. Анализ современных аналитических комплексов, основанных на технологии БЗ и обеспечивающих обработку на единой основе разнородной информации.
2. Исследование структур биллингов телефонных переговоров и банковских счетов с целью создания унифицированного процессора для их преобразования в РСС.
3. Обеспечение возможности совместного использования структур знаний в виде РСС, представляющих тексты на ЕЯ и биллинги, для решения существующих задач Системы.
4. Изучение специальных задач пользователей, основанных на информации о телефонных переговорах и банковских переводах, для разработки новых средств их решения в рамках Системы.
5. Разработка новых логико-аналитических режимов обработки информации, представленной в виде структур знаний, в рамках Системы.
6. Исследование особенностей представления информации в БЗ и разработка методов представления структур знаний в СУБД Oracle для повышения эффективности хранилища знаний Системы.
7. Исследование информационных процессов, связанных с задачей интеграции данных, и разработка технологии интеграции БЗ Системы с внешними БД для расширения пространства поиска Системы.

Методы исследования. Для решения поставленных задач в диссертации использовались методы математической логики, методы обработки структур знаний, методы формальных грамматик, методы многомерного статистического анализа (кластерный, частотный и временной анализы).

Научная новизна. В работе получены следующие новые научные результаты:

1. Проведено исследование и сравнительный анализ систем, основанных на технологии БЗ и обеспечивающих обработку на единой основе разнородной информации.
2. Разработаны алгоритмы извлечения знаний из биллингов различных структур (телефонных переговоров, банковских переводов) и их отображения в БЗ.
3. Разработана методика аналитической обработки биллингов на основе информации в БЗ (решение задачи группировки связанных объектов с учетом частоты их появления, а также их визуализация в виде временных гистограмм и графов).
4. Разработан метод анализа временных совпадений, обеспечивающий на основе информации в БЗ выявление временной связи между интересующими пользователя событиями.
5. Проведено исследование специфики применения методов кластерного анализа к биллингам телефонных переговоров. Выявлена комбинация метрик и алгоритмов кластерного анализа, позволяющая осуществить оптимальное разбиение телефонов на кластеры.
6. Предложена методика инкапсуляции структур знаний в реляционную СУБД, позволяющая обеспечить работу Системы с большими объемами данных.
7. Разработана методика интеграции Системы, основанной на структурах знаний, с внешними БД.

Достоверность научных положений, рекомендаций и выводов. Обоснованность научных положений, рекомендаций и выводов определяется корректным использованием математических методов и моделей.

Достоверность положений и выводов диссертации подтверждена результатами исследований и экспериментальными данными, полученными при внедрении Системы. Предложенные определения и классификации апробированы на конференциях и в научных публикациях.

Практическая значимость. В работе получены следующие практически значимые результаты:

1. Разработаны программные компоненты, обеспечивающие в рамках Системы обработку новых источников информации, анализ накопленной информации и оптимизацию работы хранилища знаний.
2. Разработана информационная технология комплексной обработки разнородной информации, которая может служить основой для создания новых программных систем, ориентированных на решение сложных логико-аналитических задач в различных предметных областях.

Реализация результатов работы. Результаты представлены:

1. В двух научно-исследовательских отчетах ИПИ РАН – № гос. регистрации 0120.0 412404, № гос. регистрации 0120.0 603386 за 2005, 2007 гг.

2. В программе "Логико-аналитическая система «Криминал»", внедренной в ГУВД города Москвы в 2002-2004 гг. в рамках договора № 61-И/01 (992-14-И) между Московским комитетом науки и технологии и ИПИ РАН;
3. В программе "Логико-аналитическая система обработки документов «Аналитик»" – свидетельство РОСПАТЕНТа № 2006610239 от 10.01.2006 г.
4. В учебном процессе Московского университета МВД России.

Апробация работы. Основные положения диссертационной работы докладывались и обсуждались на международной научной конференции MegaLing'2007 «Горизонты прикладной лингвистики и лингвистических технологий» (Партенит, 2007), на II Научной сессии ИПИ РАН «Проблемы и методы информатики» (Москва, 2005), на научно-технической конференции кафедры «Системы обработки информации и управления» МГТУ им. Н.Э.Баумана (Москва, 2002), на научно-технической конференции МТУСИ (Москва, 2002).

Публикации. По тематике диссертационной работы имеется 12 печатных публикаций, в том числе две в рекомендованных ВАК журналах. Кроме того, по теме диссертации опубликованы материалы в двух научно-технических отчетах ИПИ РАН за 2005 и 2007 гг.

Структура работы. Диссертация состоит из введения, четырех глав, заключения и двух приложений. Содержание работы изложено на 194 страницах, иллюстрированных 24 таблицами и 66 рисунками. Список использованных источников содержит 139 наименований.

КРАТКОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во введении обосновывается актуальность работы. Ставятся основные цели и задачи исследования.

Первая глава состоит из двух параграфов. *В первом параграфе* для определения базового функционала Системы на примере системы «Аналитик» проводится исследование особенностей и компонентов типовой системы, основанной на технологии БЗ. Анализируются основные задачи, решаемые этой системой: автоматический ввод документов с их делением на части и лексическим анализом; автоматическая формализация текстовой информации с созданием собственной БЗ – имеется в виду направленное извлечение знаний с помощью лингвистического процессора (ЛП) из текстов на ЕЯ с их использованием на уровне БЗ и др.

В качестве структур знаний в БЗ выступают расширенные семантические сети, зарекомендовавшие себя как эффективное средство представления знаний, позволяющее отобразить особенности ЕЯ. РСС состоят из однотипных N-арных фрагментов. В каждый из них введена вершина, называемая кодом фрагмента и соответствующая всей представленной в нем информации. Помимо этого вводится множество "внутренних" вершин, которые порождает сама Система по мере необходимости и которые сопоставляются неименованным объектам. Наличие этих вершин обеспечивает достаточную

универсальность РСС и позволяет сохранить семантические компоненты текстов на ЕЯ. По этим же причинам логические возможности РСС выходят за рамки возможностей логики предикатов 1-го и 2-го порядков. На основе РСС в системе «Аналитик» осуществляется вся аналитическая обработка и поиск.

При использовании технологии БЗ, в отличие от БД, структуры знаний не ограничиваются какими-либо схемами. Любая глагольная форма с ее обязательными и факультативными актантами навязывает свою схему, которая представляется в виде фрагмента РСС (рис. 1, 2).

Багдад. 9.05.06. ИНТЕРФАКС/АФП. 8 июня был совершен акт саботажа на нефтепроводе на севере Ирака …

Рисунок 1. Текст СМИ

```
PLACE_(ГОРОД,БАГДАД/1+)
ДАТА_(2006,ИЮНЬ,9/2+)
ОРГ_(ИНТЕРФАКС,АФП/3+)
СОВЕРШИТЬ(АКТ,САБОТАЖ,НА,НЕФТЕПРОВОД/4+)
ДАТА_(8,ИЮНЬ/5+)
Когда(4-,5-)
PLACE_(СЕВЕР,ГОС-ВО,ИРАК/6+)
Где(4-,6-)
```

Рисунок 2. РСС текста СМИ

На рисунке 2 представлено, что действие "СОВЕРШИТЬ" (с кодом 4+, 4-) связано с датой происшествия "ДАТА_" (с кодом 5+, 5-) через фрагмент "Когда" и с местом происшествия "PLACE_" (с кодом 6+, 6-) через фрагмент "Где". Упомянутое действие (с кодом 4+, 4-) может быть связано с другими действиями. Такие структуры формируются автоматически с помощью семантико-ориентированного ЛП, основанного на структурах знаний. Он обеспечивает автоматическое построение по текстам на ЕЯ их содержательных портретов в виде РСС.

Логико-аналитическая обработка осуществляется на основе РСС программами, написанными на языке Декл. В результате удается решать новые задачи, связанные с семантическим поиском, экспертными оценками, принятием оперативных решений. Например, в системе «Криминал», разработанной на основе системы «Аналитик», решаются следующие наукоемкие задачи:

- поиск похожих происшествий;
- поиск похожих фигурантов по приметам;
- поиск информации фактографического характера по запросам на ЕЯ и др.

В то же время в системе «Аналитик» имеется ряд проблем. Во-первых, для хранения структур знаний в системе «Аналитик» используется своя внутренняя база данных, основанная на плоских файлах. Учитывая объемы существующих потоков данных, возникает необходимость использовать в качестве хранилища знаний современные СУБД (например, Oracle, MSSQL), обеспечивающие работу с большими объемами информации.

Во-вторых, не реализовано взаимодействие с внешними источниками данных: телефонными справочниками, адресными книгами и другими данными, введенными в соответствующие БД (например, "Кронос", ГИБДД) и широко используемыми в криминальной милиции. Таким образом, необходимо организовать эффективное взаимодействие внешних БД с БЗ Системы.

В-третьих, одна из проблем связана с аналитической обработкой слабоструктурированной информации – биллингов телефонных переговоров и банковских переводов. В зависимости от компании-автора биллинги могут иметь различную структуру. Возникает задача разработки интегрированного универсального средства извлечения и представления в БЗ информации из биллингов, а также логико-аналитических режимов для ее анализа.

Эти проблемы решаются в рамках предлагаемой Системы.

Во втором параграфе проводится исследование двадцати отечественных и зарубежных аналогов системы «Аналитик»: «SynSys Semantix», «Интегрум», «RetrievalWare» и др. Определяются современные перспективные методы обработки разнородной информации, которые могли бы быть реализованы в Системе. Проводится выбор параметров сравнения этих систем и их обоснование. Представлена таблица сравнения аналогов в разрезе выбранных параметров. По результатам сравнения формируются требования и предполагаемый состав функций, которые должны быть реализованы в Системе. В конце первой главы сделаны выводы, в которых предлагается разработать новую информационную технологию комплексной обработки разнородной информации на основе технологических решений системы «Аналитик». Технология должна быть основана:

1) на расширении информационного пространства Системы, посредством подключения к Системе в качестве исходных данных слабоструктурированной информации (биллингов телефонных переговоров и банковских переводов) и внешних БД;
2) на расширении аналитических возможностей Системы, путем реализации новых логико-аналитических режимов обработки накопленной информации на основе структур знаний в виде РСС;
3) на применении современных технологий в области хранения информации.

В первом параграфе **второй главы** описывается развитие аналитических возможностей Системы. Для анализа неструктурированной информации (текстов на ЕЯ) разработан режим «Анализ временных совпадений». Этот автоматический режим позволяет аналитику на основе информации из БЗ выявить связанные происшествия, произошедшие в один и тот же период времени. При больших объемах информации выявить подобные связи вручную человеку практически невозможно.

Вся обработка в этом режиме проходит на уровне РСС. В процессе поиска документов в БЗ по выделенным в исходном документе объектам используется режим «Поиск похожих», где всем полученным в результате поиска документам в зависимости от степени совпадения найденной и исходной информации присваивается тот или иной вес. Результаты поиска с наибольшим

весом сравниваются с исходным документом на предмет совпадения по времени и, если временной интервал совпадает, визуализируются. Для визуализации разработан специальный модуль, отображающий результаты анализа в виде блочных структур на временной оси (рис. 3).

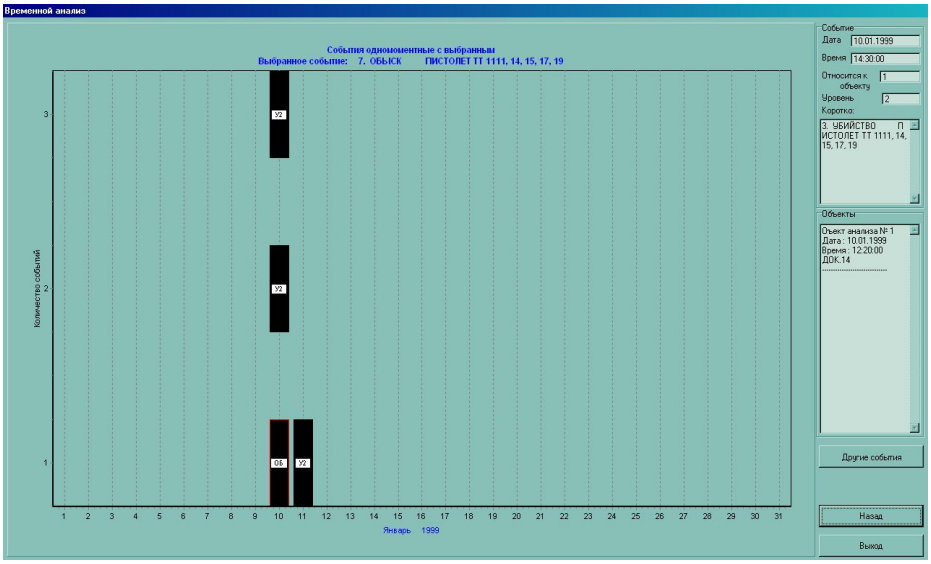

Рисунок 3. Визуализация событий в режиме «Анализ временных совпадений»

Страницы 5 из 8 Номер SIM-карты:8970101000000654321
Телефон номер 2234567 Сведения о разговорах:
Дата Время Номер Зона ПС Зона ВТК Услуга Длит. Стоимость
01/07/00 12:17 6624128 Телеф 32:07 0.52
01/07/00 13:57 <-0956624128 Телеф 1:59 0.52

Рисунок 4. Фрагмент биллинга 1

Звонки с номера 0959222301 за интервал с 01.01.2002 по 06.08.2002
imsi - Амирас Азер Агалиоглы лиц. счет - 1746762 Частное лицо
Mob_num Num Dur Dir Num_A Cell_Id Bts_addr
0959222301 +70951079565 01.01.02 13:30:26 168 O Неизвестно
0959222301 +70951713495 01.01.02 14:55:12 3 O Неизвестно

Рисунок 5. Фрагмент биллинга 2

Развитие второго направления технологии – расширение информационного пространства – осуществляется за счет подключения к Системе в качестве источника слабоструктурированной информации биллингов (рис. 4, 5). Этот вид табличных данных существует в таких областях, как телекоммуникации или банковский сектор. В телекоммуникационном секторе биллинги представляют собой расшифровку всех телефонных разговоров, SMS сообщений и прочих платных услуг, сделанных с определенного телефонного номера, и обычно прикладываются к счету за услуги связи. В банковском секторе биллинг – это расшифровка всех денежных переводов, сделанных с/на определенный счет. На сегодняшний день практически у каждого городского жителя есть телефон или банковский счет (банковская карточка), по каждому из которых в компании-поставщике услуг делается ежемесячная детализация (биллинг) всех его разговоров или денежных переводов. Таким образом, речь

идет о миллионах документов. Из приведенных примеров (рис. 4, 5) видно, что структура биллинга телефонных переговоров в зависимости от компании-автора (МТС, Билайн, Мегафон и т.д.) различна (с банковскими переводами ситуация точно такая же).

После анализа 10 различных форматов биллингов телефонных переговоров была выявлена типовая структура биллинга, состоящая из заголовочной и основной части. В *заголовочной части* были выделены 11 атрибутов, содержащих информацию об объекте детализации: телефон, ФИО, адрес, период детализации, номер SIM-карты и пр. В *основной части* биллинга находится информация, описывающая повторяющиеся во времени события: звонки, SMS и MMS сообщения, GPRS пакеты и т.п. Здесь было выделено 25 различных атрибутов. В заголовочной части Z биллинга выделены следующие основные атрибуты:

$$Z\ (N_1,\ D_1,\ D_2,\ F_1),$$

где: N_1 – телефонный номер детализации;
D_1 – дата начала детализации;
D_2 – дата окончания периода детализации;
F_1 – ФИО лица, которому принадлежит телефонный номер детализации.

В *основной части* выделяются следующие атрибуты, используемые в аналитической обработке, которые повторяются практически во всех биллингах телефонных переговоров: дата соединения, время соединения, "Телефон А", "Телефон Б", длительность, стоимость.

Строка S биллинга описывается следующими атрибутами:

$$S\ (\ Num_1,\ D_1,\ Dlit,\ Nap),$$

где: $Num_1 = \{n_1, n_2, \ldots, n_r\}$ – все неповторяющиеся номера телефонов детализации, на/с которых произошло соединение с N_1, r – количество неповторяющихся номеров телефонов детализации;
D_1 – дата соединения;
$Dlit = \{dlit_1, dlit_2, \ldots, dlit_f\}$ – длительность соединения, f – количество строк в детализации;
$Nap = \{nap_1, nap_2, \ldots, nap_f\} = \{исх, вх\}$ – направление соединения (исходящее или входящее соединение).

В работе предложены и апробированы методы распознавания биллингов различных структур, на основе которых разработан семантический анализатор, представляющий собой интегрированное средство извлечения и преобразования в РСС находящейся в биллингах информации. Извлечение знаний осуществляется при помощи разрабатываемого в визуальной среде шаблона распознавания и применения набора контекстных правил.

Для логико-аналитической обработки биллингов в Системе реализован режим «Детализация номерных объектов», который можно условно разделить на четыре подрежима: «Граф телефонных переговоров», «Диаграмма длительности переговоров», «Граф финансовых потоков», «Диаграмма финансовых потоков». Режимы «Граф телефонных переговоров» и «Граф

финансовых потоков» решают задачу классификации. Они позволяют аналитику автоматически из всего массива информации (в биллинге за год может быть более десяти тысяч соединений) выявить наиболее активные телефонные номера или счета. Режимы «Диаграмма длительности переговоров» и «Диаграмма финансовых потоков» позволяют выявить пики активности в работе телефона или счета, на который сделана детализация. В криминалистике, например, это позволяет только с помощью детализации определить время подготовки преступления. В зависимости от режима анализа информация визуализируется двумя видами: графом и диаграммой (рис. 6).

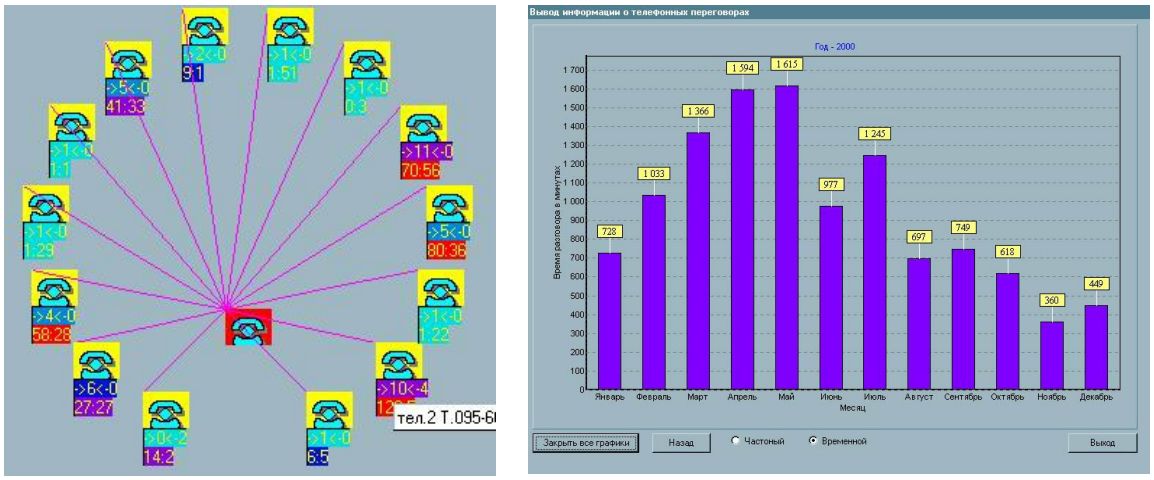

Рисунок 6. Визуализация биллингов телефонных переговоров

Вся логико-аналитическая часть режимов реализована на языке Декл и происходит на уровне структур знаний в виде РСС. На основе анализа, проведенного по всей БЗ, выделяются все биллинги по выбранному номеру – объекту исследования (ОИ). Таким образом, у аналитика есть возможность анализировать активность ОИ за любые промежутки времени, за которые по нему в БЗ есть биллинги. Проводится определение особо активных телефонов, с которыми ОИ разговаривал либо чаще всего, либо дольше всего. Т.е. ведется подсчет количества входящих и исходящих звонков между ОИ и другими телефонами за выбранный промежуток времени и подсчет длительности переговоров.

Результатом работы подрежима «Граф телефонных переговоров» является применение следующих правил. Для каждой строки детализации $m=\{1..f\}$ и каждого телефонного номера n_p из Num_1, где $p=\{1..r\}$: при условии, что $nap_m = исх$ вычисляется $S_{1p} = \sum_{i=1}^{k_p} dlit_m$, где k_p – количество строк детализации, удовлетворяющих этим условиям; при условии, что $nap_m = вх$ вычисляется $S_{2p} = \sum_{i=1}^{l_p} dlit_m$, где l_p – количество строк детализации, удовлетворяющих этому условию. В результате получается следующий набор атрибутов результирующей агрегированной информации *Itog*:

$Itog\ (Num_1, S_1, S_2, K_1, K_2)$, где:

$Num_1 = \{n_1, n_2, ..., n_r\}$ – множество неповторяющихся номеров телефонов;
$S_1 = \{ S_{11}, S_{12}, ..., S_{1p} \}$ – сумма длительностей исходящих соединений;

$S_2 = \{ S_{21}, S_{22}, \ldots, S_{2p} \}$ – сумма длительностей входящих соединений;
$K_1 = \{k_1, k_2, \ldots, k_p\}$ – количество исходящих соединений;
$K_2 = \{l_1, l_2, \ldots, l_p \}$ – количество входящих соединений, $p = \{1..r\}$.

В каждом из режимов визуализации реализован дополнительный функционал. Для графа – это различная раскраска объектов в зависимости от их активности, отображение графа в виде дерева или окружности. Для диаграммы – это визуализация в различных масштабах (год, месяц, день), возможность ввода "пороговых дат" (дней, когда произошли какие-то важные события), возможность сравнения активности ОИ в течение двух различных дней (эта функция необходима для выявления необычной активности ОИ). Обработка и визуализация банковских счетов аналогична работе с биллингами телефонных переговоров, но имеет некоторые особенности в графическом отображении.

Во втором параграфе описан следующий этап развития разрабатываемой технологии в направлении логико-аналитической обработки – проводится исследование моделей и алгоритмов кластерного анализа. С помощью этого типа анализа решается задача классификации объектов по множеству атрибутов. Результаты исследования могут использоваться в решении следующих задач: выявление преступных групп лиц и связей в криминалистике, разработка новых тарифов в мобильной связи, выявление групп счетов в банковском секторе и др.

Результат кластеризации зависит от ряда факторов, таких как разнородность данных, наличие выбросов, наличие взаимосвязанных объектов, предметной области, объемов обрабатываемых данных и т.п. Для достижения лучшего результата необходимо найти оптимальные критерии кластеризации биллингов. Для определения этих критериев проводится следующий эксперимент. В качестве исходных данных рассмотрен биллинг с деперсонифицированными данными за месячный период с общей продолжительностью соединений 226 часов, состоящий из 4704 строк. Эти данные с помощью последовательности группировок "вручную" разбиваются на оптимальное с точки зрения аналитика число кластеров. После этого данные перемешиваются между собой так, чтобы максимально усложнить задачу автоматической кластеризации. К этим данным применяются различные комбинации метрик и алгоритмов кластеризации. Оптимальной для автоматической кластеризации биллингов является та комбинация, результатом работы которой является разбиение, максимально похожее на разбиение "вручную".

Исходный биллинг с помощью последовательности группировок был разбит на 10 групп телефонов – экземпляров *Itog* (табл. 1). После расчета средней длительности входящих и исходящих вызовов по формулам $d_{вх} = S_2 / K_2$ и $d_{исх} = S_1 / K_1$ выделены 4 группы-кластера телефонов Num_1:

1) π_1 = {111 (121, 1200, 5, 112), 555 (154, 978, 8, 88), 888 (113, 1144, 1, 70)};
2) π_2 = {222 (878, 200, 99, 4), 444 (500, 300, 60, 12), 777 (1400, 400, 130, 20)};
3) π_3 = {666 (2500, 232, 2314, 29), 999 (2134, 122, 1700, 18)};
4) π_4 = {333 (21, 600, 1, 10), 900 (11, 578, 7, 16)}.

Таблица 1.

Исходные данные кластеризации

Номер телефона (Num*1*)	**Длительность исходящих соединений (*S1*)**	**Длительность входящих соединений (*S2*)**	**Количество исходящих соединений (*K1*)**	**Количество входящих соединений (*K2*)**
111	121	1200	5	112
222	878	200	99	4
333	21	600	1	10
444	500	300	60	12
555	154	978	8	88
666	2500	232	2314	29
777	1400	400	130	20
888	113	1144	1	70
999	2134	122	1700	18
900	11	578	7	16

Правильность такого разбиения подтверждается при подсчете усредненных показателей групп – средних длительностей исходящих и входящих соединений, средних количеств исходящих и входящих соединений по формулам:

$\overline{d}_{исх} = \frac{\sum_{i=1}^{n} S_{1i}}{\sum_{i=1}^{n} k_i}$, $\overline{d}_{вх} = \frac{\sum_{i=1}^{n} S_{2i}}{\sum_{i=1}^{n} l_i}$, $\overline{k}_{исх} = \frac{\sum_{i=1}^{n} k_i}{n}$, $\overline{k}_{вх} = \frac{\sum_{i=1}^{n} l_i}{n}$, где: n – количество элементов в группе, i – номер элемента в группе. Усредненные показатели кластеров представлены в таблице 2.

Таблица 2.

Усредненные показатели групп

Номер группы	**Средняя длительность исходящих по группе ($\overline{d}_{исх}$)**	**Средняя длительность входящих по группе ($\overline{d}_{вх}$)**	**Среднее количество исходящих ($\overline{k}_{исх}$)**	**Среднее количество входящих ($\overline{k}_{вх}$)**
1	52,15	12,7	4,7	90
2	27,96	31,7	96,3	12
3	1,17	7,4	2007	23,5
4	11,25	48	4	13

Для оценки результатов кластеризации используется целевая функция – сумма квадратов отклонений (СКО). СКО рассчитывается по формуле $W = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} (\sum_{i=1}^{n} x_i)^2$. Для разбиения "вручную" *W=721475.3.*

Для более точного определения оптимальных параметров разбиения автоматическая кластеризация осуществляется с помощью различных комбинаций метрик и методов кластеризации в различных программных комплексах: SPSS, Statgraphics (STAT.) и Attestat (табл. 3).

Таблица 3.
Сводная таблица кластеризации

Номер группы / **Номер телефона**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
111	*1*	*1*	*1*	*1*	1	1	1	1	1
222	*2*	*2*	*2*	*2*	4	4	2	2	2
333	*3*	*3*	*3*	*3*	1	1	3	1	1
444	*2*	*2*	*2*	*2*	4	4	2	2	2
555	*1*	*1*	*1*	*1*	1	1	1	1	1
666	*4*	*4*	*4*	*4*	2	3	4	3	3
777	*2*	*2*	*2*	*2*	4	3	3	2	2
888	*1*	*1*	*1*	*1*	1	1	1	1	1
999	*4*	*4*	*4*	*4*	3	3	4	4	4
900	*3*	*3*	*3*	*3*	1	1	3	1	1
Метод	Уорд	Уорд	Уорд	k-средних	k-средних	k-средних	k-средних	медианы	средней связи
Метрика	сити-блок, минковского	Махаланобис, эвклидова, сити-блок	сити-блок, эвклидова	сити-блок	эвклидова	эвклидова	эвклидова	ближнего соседа, дальнего соседа, центроидный, сити-блок	сити-блок
Программа	SPSS	Attestat	STAT.	STAT.	SPSS	Attestat	STAT.	SPSS	Attestat, SPSS
СКО	*721476*	*721476*	*721476*	*721476*	804821	3655103	1679066	804820.6	804821

В первых десяти строках таблицы 3 представлены объекты и результаты кластеризации – номера кластеров, в которые был помещен тот или иной телефонный номер. Курсивом выделены результаты, совпадающие с результатами разбиения "вручную". В нижней части таблицы представлены методы кластеризации (использовались как иерархические методы, так и неиерархические), метрики и программы кластеризации. В качестве исходных данных задавалось количество кластеров *m=4*, на которое разбивалось множество объектов кластеризации.

На основании полученного СКО можно сделать следующий вывод: более точные результат дает сочетание частных случаев метрики Минковского (сити-блок, эвклидова метрика) и алгоритмов иерархической кластеризации (метод Уорда), что позволяет рекомендовать их использование при кластеризации детализаций телефонных переговоров.

Рисунок 7. Фрагмент схемы БД в СУБД Oracle

В первом параграфе **третьей главы** решается задача оптимизации работы хранилища знаний. В системе «Аналитик» БЗ хранится в БД на плоских файлах. Такой способ хранения обладает рядом недостатков: медленная обработка при больших объемах данных, высокая трудоемкость удаления, сложность поиска и др. Помимо этого возникают проблемы защиты данных и управления. В то же время в современных СУБД, таких как Oracle, MSSQL, MySQL эти проблемы уже решены. Проводится анализ требований к СУБД, на основе которой разрабатывается хранилище знаний Системы. В качестве СУБД, удовлетворяющей всем предъявленным требованиям, выбирается Oracle. Предлагается метод, обеспечивающий хранение структур знаний Системы в этой СУБД. Проектируется новая схема БД в Oracle, позволяющая хранить РСС (рис. 7), что решает вышеперечисленные проблемы в рамках предлагаемой Системы. Приводятся примеры реализации удаления/изменения документов, поиска и способов обеспечения безопасности при использовании новой схемы хранения знаний.

Во втором параграфе описана методика подключения к БЗ внешних источников данных – еще один этап развития технологии в направлении

расширения информационного пространства Системы. В качестве внешних источников данных предлагается использовать специализированные базы данных, например, базы МГТС или ГИБДД. Проводится анализ современных методов интеграции данных: промышленных средств интеграции данных (ПСИД), сервисных шин, интеграции с помощью "point-to-point" интерфейсов и адаптеров. ПСИД и сервисные шины являются очень дорогим решением. Разработка "point-to-point" интерфейсов требует участия специалистов по программированию в области баз данных в процессе эксплуатации Системы. Адаптеры в качестве канонической модели данных используют язык Синтез, а не РСС. В результате предлагается новая методика интеграции внешних баз данных с БЗ Системы на основе редактора шаблонов соединений, не требующая от пользователя каких-либо специальных навыков. В работе представлен пример взаимодействия Системы с базой МГТС при помощи предложенной методики. Найденная во внешней базе данных информация может пополнять собственную БЗ Системы.

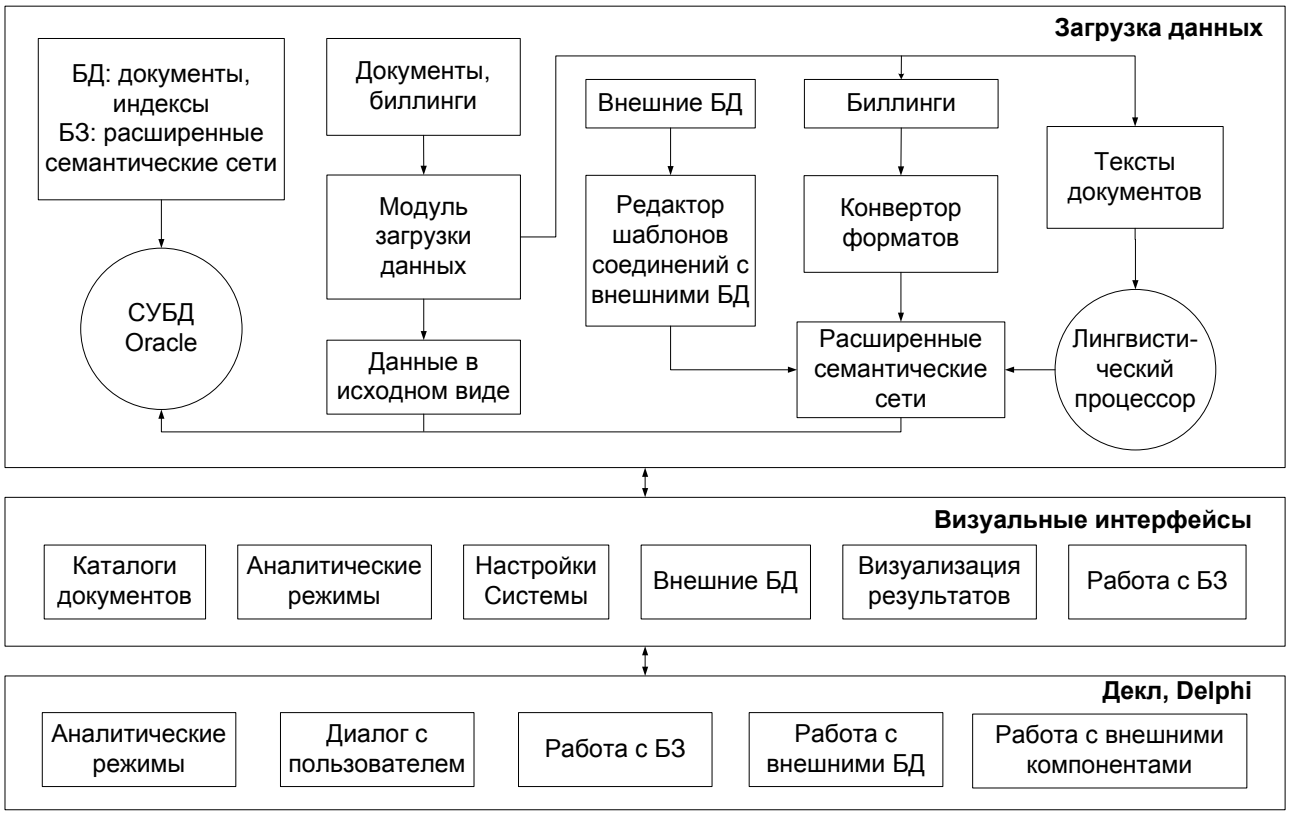

Рисунок 8. Структурная схема Системы

В результате в рамках диссертационной работы разработана интегрированная информационная технология комплексной обработки информации на основе структур знаний в виде РСС. Технология включает в себя: этапы автоматической обработки потоков разнородной информации, существующей в электронном виде; методы загрузки этой информации в хранилище знаний; методы и алгоритмы логико-аналитической обработки накопленной информации; средства визуализации результатов обработки; методы поиска и извлечения информации из внешних баз данных.

В **четвертой главе** описывается программная реализация интегрированной информационной технологии. Представлены: структурная схема Системы (рис. 8); модульная структура программы, состоящая из 24 разработанных в рамках диссертационной работы модулей; граф диалога пользователя; модули распознавания биллингов телефонных переговоров и банковских переводов; модули ручного ввода в Систему и их преобразования в РСС биллингов телефонных переговоров и записных книжек с контактами абонентов. Описаны процедуры создания схемы БД в Oracle, процедуры обращения к СУБД из Системы, интерфейс взаимодействия пользователя с Системой.

В заключении приводятся основные выводы, полученные в работе.

В приложения вынесены поясняющие и вспомогательные материалы.

ОСНОВНЫЕ ВЫВОДЫ И РЕЗУЛЬТАТЫ РАБОТЫ.

1. Разработана новая информационная технология комплексной обработки разнородной информации большого объема в рамках Системы, основанной на структурах знаний в виде расширенных семантических сетей.
2. По итогам проведенного исследования систем, основанных на технологии баз знаний, в качестве единого средства представления разнородной информации (текстов на естественном языке, биллингов, данных из внешних баз) предложено использовать расширенные семантические сети.
3. На основе исследования структур биллингов разработан семантический анализатор – интегрированное средство извлечения данных из биллингов и их представления в виде расширенных семантических сетей.
4. Разработаны методика и алгоритмы решения задачи детализации номерных объектов, позволяющие группировать связанные объекты (телефонные номера, банковские счета) на основе информации из базы знаний.
5. Впервые проведено исследование специфики применения кластерного анализа к биллингам телефонных переговоров. Выявлена комбинация метрики и алгоритма кластерного анализа, позволяющая осуществить оптимальную с точки зрения целевой функции кластеризацию.
6. Разработан режим «Анализ временных совпадений», позволяющий аналитику увидеть временную связь между интересующими его событиями.
7. Предложена методика инкапсуляции структур знаний в реляционную СУБД, что позволяет обеспечить работу Системы с большими объемами данных.
8. Для расширения пространства поиска разработана методика интеграции базы знаний Системы с внешними базами данных на основе редактора шаблонов соединений.
9. Разработана программная реализация предложенной технологии.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИОННОЙ РАБОТЫ.

1. Рабинович Б.И. Редактор шаблонов соединений как средство интеграции базы знаний системы «Аналитик» с внешними источниками данных // Вестник МГТУ им. Н.Э. Баумана. Серия «Приборостроение». №2. – М.: МГТУ им. Н.Э. Баумана, 2008. – С. 113-121.

2. Рабинович Б.И. Обзор информационных систем анализа текстов на естественном языке // Известия высших учебных заведений. Проблемы полиграфии и издательского дела. №2. – М.: МГУП, 2008. – С. 83-88.
3. Рабинович Б.И. Электронное хранилище разнородной информации на основе структур знаний // Информатизация и связь. Специальный выпуск'2008. – М.: «Информатизация и связь», 2008. – С. 84-88.
4. Кузнецов И.П. Рабинович Б.И. Модель базы знаний с возможностью интеграции внешних источников информации в системе «Аналитик» // Системы и средства информатики. Ин-т пробл. информатики РАН. Вып. 17 / Отв. ред. И.А. Соколов. – М.: Наука, 2007. – С. 254-272.
5. Рабинович Б.И. Кластерный анализ детализаций телефонных переговоров // Системы и средства информатики. Ин-т пробл. информатики РАН. Вып. 17 / Отв. ред. И.А. Соколов. – М.: Наука, 2007. – С. 52-78.
6. Рабинович Б.И. Система обработки потоков данных // MegaLing'2007 Горизонты прикладной лингвистики и лингвистических технологий Доклады международной научной конференции. 24-28 сентября 2007, Украина, Крым, Партенит. – Симферополь: «ДиАйПи», 2007. – С. 331-332.
7. Рабинович Б.И. Организация баз знаний в современных СУБД // Проблемы и методы информатики. II Научная сессия ИПИ РАН. Москва, 18-22 апреля 2005 г. Тезисы докладов. – М.: ИПИ РАН, 2005. – С. 165-168.
8. Рабинович Б.И. Система сбора и обработки разнородной информации «Аналитик» // Интеллектуальные технологии и системы. Сборник учебно-методических работ и статей аспирантов и студентов. Выпуск 7. – М.: «Эликс+», 2005. – С. 211-230.
9. Рабинович Б.И. Хранение БЗ в современных СУБД // Интеллектуальные технологии и системы. Сборник учебно-методических работ и статей аспирантов и студентов. Выпуск 6. – М.: «Эликс+», 2004. – С. 173-186.
10. Рабинович Б.И. Аналитическая система обработки и управления структурированной информацией // Интеллектуальные технологии и системы. Сборник учебно-методических работ и статей аспирантов и студентов. Выпуск 5. – М.: «Эликс+», 2003. – С. 284-296.
11. Кузнецов И.П., Мацкевич А.Г., Рабинович Б.И., Гнидо Е.И. Временной анализ потоков событий в Логико-Аналитической системе «Аналитик» // Тезисы докладов НТК МТУСИ. 29-31 января 2002 г. – М.: Инсвязьиздат, 2002. – С. 409-410.
12. Кузнецов И.П., Мацкевич А.Г., Рабинович Б.И., Гнидо Е.И. Частотный анализ биллингов телефонных переговоров в Логико-Аналитической системе «Аналитик» // Тезисы докладов НТК МТУСИ. 29-31 января 2002 г. – М.: Инсвязьиздат, 2002. – С. 409.