На правах рукописи

КОРНЕВ ПАВЕЛ АЛЕКСАНДРОВИЧ

АЛГОРИТМЫ КАТЕГОРИРОВАНИЯ ПЕРСОНАЛЬНЫХ ДАННЫХ ДЛЯ СИСТЕМ АВТОМАТИЗИРОВАННОГО ПРОЕКТИРОВАНИЯ БАЗ ДАННЫХ ИНФОРМАЦИОННЫХ СИСТЕМ

Специальность 05.13.12 –

Системы автоматизации проектирования (технические системы)

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Липецк 2012

Работа выполнена на кафедре электроники телекоммуникаций и компьютерных технологий федерального государственного бюджетного образовательного учреждения высшего профессионального образования "Липецкий государственный педагогический университет" (г. Липецк).

Научный руководитель:	доктор технических наук, профессор МАЛЫШ Владимир Николаевич
Официальные оппоненты:	БЕЛОВ Владимир Викторович, доктор технических наук, профессор, профессор кафедры вычислительной и прикладной математики ФГБОУ ВПО «Рязанский государственный радиотехнический университет»
	КОНДРАШИН Юрий Алексеевич, кандидат технических наук, доцент, доцент кафедры информатики ФГБОУ ВПО «Липецкий государственный педагогический университет»
Ведущая организация:	ФГБОУ ВПО «Липецкий государственный технический университет» (г. Липецк)

Защита диссертации состоится 30 мая 2012 г. в 12 часов на заседании диссертационного совета Д 212.211.02 в ФГБОУ ВПО «Рязанский государственный радиотехнический университет» по адресу: 390005, г. Рязань, ул. Гагарина, 59/1.

С диссертацией можно ознакомиться в библиотеке ФГБОУ ВПО «Рязанский государственный радиотехнический университет».

Автореферат разослан «__» _____апреля_____ 2012 г.

Ученый секретарь
диссертационного совета А.И. Таганов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. В настоящее время проектирование баз данных для современных информационных систем происходит на основе различных типов САПР (систем автоматизированного проектирования). При этом САПР автоматизируют лишь рутинные действия разработчика, а основную интеллектуальную работу выполняет человек. Автор-разработчик информационной системы вынужден вручную выполнять синтез и анализ проектных решений баз данных (БД), проверяя их характеристики (защищенность, кроссплатформенность, возможность экспорта, импорта данных и т.д.) на соответствие действующим законам и нормативно-правовым актам. В частности, в связи с введением в действие обязательных положений закона "О персональных данных" (N 152-ФЗ в редакции от 25.07.2011) разработчику при проектировании типовой информационной системы надлежит обеспечить защиту определенных категорий персональных данных (ПДн). Для автоматизации процедуры анализа проектных решений баз данных на наличие блоков ПДн необходимо специальное алгоритмическое и программное обеспечение, которое на данный момент отсутствует.

Следует отметить, что наименование различных блоков ПДн (таблиц, кортежей и т.п.) является исключительно субъективной прерогативой разработчика. Поэтому нет объективной возможности для создания системы четких шаблонов, на основе которых происходил бы анализ проектных решений. Существующие в данном процессе неполнота и неточность информации, позволяют сделать вывод о необходимости применения аппарата теории нечетких множеств (ТНМ) для формализации процессов анализа проектных решений. Кроме этого, потенциальная изменчивость отечественного законодательства и нормативно-правовых актов делают перспективным дополнительное применение аппарата искусственных нейронных сетей (ИНС) в качестве "обучаемой в нестандартной ситуации" базы правил и эффективного блока вывода для систем нечеткого вывода.

Многие зарубежные и отечественные ученые ранее уделяли большое внимание прикладному значению теории нечетких множеств и нейросетевых технологий в своих научных работах, посвященных как информационной безопасности, так и другим аспектам информационной отрасли знаний (Заде Л.А., Такаги Т., Сугено М., Пилиньский М., Рутковская Д., Кудинов Ю.И., Сара-

ев П.В., Зайченко Д.Н., Полевой Н.Ю., Волков А.В., Магола Д.С., Панфилов Д.С., Хафизов А.Ф., Абрамов Е.С., Костин А.А., Свечников Л.А., Иванов А.И.).

Таким образом, актуальность настоящей работы определяется необходимостью разработки эффективных алгоритмов категорирования персональных данных на основе ТНМ и ИНС с целью повышения качества проектирования баз данных с учетом требований защиты персональных данных.

Целью диссертационного исследования является повышение эффективности проектирования баз данных информационных систем с учетом требований защиты персональных данных.

Объектом исследования являются БД информационных систем персональных данных (БД ИСПДн).

Предметом исследований являются алгоритмы категорирования персональных данных в базах данных информационных систем.

Научные задачи, решаемые в работе:

1. Разработать новые алгоритмы для автоматизированного категорирования персональных данных в базах данных ИСПДн.
2. Разработать модификации нейросетей Кохонена для автоматизированного категорирования персональных данных в базах данных ИСПДн.
3. Разработать UML-модели для автоматизированного категорирования персональных данных в базах данных ИСПДн.
4. Создать программное обеспечение с целью автоматизированного категорирования персональных данных в базах данных ИСПДн.

Для решения поставленных задач применен следующий аппарат исследования:

- теория систем;
- теория баз данных;
- теория искусственных нейронных сетей;
- теория нечетких множеств;
- математическое моделирование.

Поставленные цели и задачи определили структуру и последовательность выполнения диссертационной работы. Диссертационная работа состоит из введения, четырех глав и заключения.

Основными научными результатами, выносимыми на защиту, являются:

1. Алгоритмы автоматизированного категорирования для анализа проектных решений БД, содержащих персональные данные.
2. Модификации нейросетей Кохонена для анализа проектных решений БД, содержащих персональные данные.
3. Математические модели нечетких систем вывода Мамдани и Сугено для анализа проектных решений БД, содержащих персональные данные.
4. UML-модели систем и подсистем анализа проектных решений БД, содержащих персональные данные.

Научная новизна работы определяется разработкой новых математических моделей нечетких систем вывода (Мамдани и Сугено) и модификаций нейросетей Кохонена и создание на их основе алгоритмов и UML-моделей для автоматизированного категорирования ПДн.

Теоретическая значимость работы заключается в обобщении теории и развитии моделей и алгоритмов анализа и синтеза проектных решений баз данных в условиях реализации информационной безопасности.

Практическая значимость работы заключается в том, что результаты исследования могут быть использованы разработчиками БД ИСПДн в совокупности с различными CASE-средствами в целях повышения эффективности проектирования БД в соответствии с современными требованиями защиты ПДн. В частности, разработанные алгоритмы и UML-модели могут быть использованы специалистами в качестве основы для создания собственных модулей и библиотек, встраиваемых в стандартные CASE-средства.

Соответствие паспорту специальности.

Согласно паспорту специальности 05.13.12 «Системы автоматизации проектирования (технические науки)», проблематика, рассмотренная в диссертации, соответствует следующим областям исследований:

"Разработка научных основ построения средств САПР, разработка и исследование моделей, алгоритмов и методов для синтеза и анализа проектных решений, включая конструкторские и технологические решения в САПР и АСТПП".

Апробация работы. По результатам исследований опубликовано 13 печатных работ.

Результаты работы были представлены на следующих международных, всероссийских и межвузовских конференциях.

1. XI Международная научно-практическая конференция "Проблемы образования в современной России и на постсоветском пространстве" (г. Пенза, 2008 г.).
2. V Всероссийская школа-семинар молодых ученых "Управление большими системами» (г. Липецк, 2008 г.).
3. IX Международная научно-методическая конференция «Информатика: проблемы, методология, технологии» (г. Воронеж, 2009 г.).
4. XXIV Международная научно-техническая конференция "Математические методы и информационные технологии в экономике, социологии и образовании" (г. Пенза, 2009 г.).
5. Межрегиональная научно-методическая конференция "Актуальные проблемы современного образования" (г. Воронеж, 2009 г.).
6. II Всероссийская научно-практическая конференция "Инновации и информационные технологии в образовании"(г. Липецк, 2010 г.).
7. III международная заочная научно-практическая конференция "Актуальные вопросы технических, экономических и гуманитарных наук" (г. Георгиевск, 15-17 июня 2010 г.).
8. XII Международная научно-техническая конференция "Информационно-вычислительные технологии и их приложения" (г. Пенза, 2010 г.).

Внедрение. Результаты диссертационного исследования адаптированы для работы с распределенными ИСПДн в рамках перспективной Intranet-сети ФГБОУ ВПО "Липецкий государственный педагогический университет". Результаты работы активно используются в курсах "Криптографическая защита информации", "Защита информационных процессов в компьютерных системах", "Экономика защиты информации". Результаты работы внедрены в производственный процесс ОАО "Грязинский культиваторный завод".

Объем работы. Общий объем работы 152 страницы машинописного текста. Список литературы включает 104 наименования.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность избранной темы и определены: задача, цель и вопросы исследования; показана научная новизна и практическая

значимость диссертации, сформулированы выносимые на защиту положения; приведены апробации результатов исследования в условиях реальных информационных систем персональных данных.

В первой главе диссертационной работы рассматриваются принципы управления информационной безопасностью персональных данных на основе автоматизации категорирования персональных данных. Произвести автоматизацию процесса категорирования невозможно без учета законодательно закрепленных норм и определений.

В частности, следует остановиться на следующих законодательно закрепленных терминах.

Персональные данные (ПДн) – это любая информация, относящаяся к определенному физическому лицу (фамилия, имя, отчество, дата и место рождения, адрес проживания, образование, доходы и т.п.).

Информационная система персональных данных (ИСПДн) – информационная система, представляющая собой совокупность персональных данных, содержащихся в базе данных, информационных технологий и технических средств.

Проведен анализ существующих мер защиты ИСПДн и исследованы основные направления автоматизации принятия управляющих решений при обеспечении информационной безопасности ПДн (рис. 1).

Следует отметить, что значительные затраты времени и сил программистов при разработке ИСПДн уходят на то, чтобы отнести содержимое таблиц разрабатываемой БД к некоторой категории ПДн (для этого требуется анализировать лингвистический смысл как имен таблиц, так и имен полей таблиц). При создании различных шаблонов с помощью регулярных выражений происходит сокращение трудо-временных затрат. При этом созданная база шаблонов статична и при её модификации или настройке на конкретную ИСПДн потребуется дополнительные затраты. Для того чтобы избежать подобных излишних трудо-временных затрат целесообразно использовать математический аппарат нечетких множеств и искусственных нейронных сетей.

Анализируя все вышеизложенное, следует сделать вывод о способах решения проблемы автоматизированного принятия управляющих решений (категорирования персональных данных), который представлен на рис. 2.

Рис. 1. Диаграмма гипотетической автоматизации принятия управляющих решений

Рис. 2. Способы автоматизации категорирования ПДн

Во второй главе рассмотрен математический аппарат, необходимый для синтеза систем автоматизированного категорирования персональных данных. Необходимо остановится на следующих основных моментах.

При защите персональных данных, содержащихся в современных информационных системах (базах данных), нет возможности четко с математической точки зрения описать как должны быть представлены персональные данные, то есть невозможно построить четкий шаблон или систему четких шаблонов. Поэтому необходимо для этого использовать математический аппарат теории нечетких множеств.

На рисунке 3 представлена открытая модель модуля нечеткого управления. Он состоит из следующих блоков:

- база правил;
- блок фаззификации;
- блок выработки решения;
- блок дефаззификации.

Лингвистическая модель (база правил) представляет собой набор нечетких правил $R^{(k)}, k = 1,...,N$, вида:

$$R^{(k)}: \mathbf{IF}\left(x_1 \text{ } это \text{ } A_1^{(k)} \ \mathbf{AND}\ x_2 \text{ } это \text{ } A_2^{(k)} ... \ \mathbf{AND}\ x_n \text{ } это \text{ } A_n^{(k)}\right) \quad (1)$$

$$\mathbf{THEN}\left(y_1 \text{ } это \text{ } B_1^k \ \mathbf{AND}\ y_2 \text{ } это \text{ } B_2^k ... \ \mathbf{AND}\ y_m \text{ } это \text{ } B_m^k\right),$$

где N – количество нечетких правил, A_i^k – нечеткие множества:

В рассматриваемом случае (двухфакторного нечеткого категорирования ПДн) обобщенный нечеткий алгоритм вывода преобразуется к следующему виду:

1. Условие: $\mathbf{X} = \left(x_1, x_2\right)^T$ это A', $A' = A_1' \times A_2'$; (2)

2. Импликация: $\bigcup_{k=1}^{12} R^{(k)}$, $R^{(k)}: A^k \to B^k$, $A^k = A_1^k \times A_2^k$;

3. Вывод: y это B'.

В ходе исследования были разработаны значения элементов множеств A_1^k, A_2^k и B^k. Следует отметить, что разрабатывались как базовые лингвистические терм-множества (с функциями принадлежностями типа синглетон), так и расширенные лингвистические терм-множества (с альтернативными функциями принадлежности).

При некоторых допущениях можно модифицировать систему нечеткого управления посредством ИНС с прямым распространением сигнала. Далее на основе экспериментальных исследований следует синтезировать архитектуру ИНС, наиболее адаптированную для кластеризации (категорирования) персональных данных.

Архитектура нейронной сети Кохонена для кластеризации ПДн (1 слой, 12 нейронов) выглядит следующим образом (рис. 4).

В нейронной сети используется конкурирующая функция активации. При этом алгоритм обучения сети Кохонена представим следующим образом:

1. Фиксируется число нейронов, начальные веса $\mathbf{w}_i(0)$ и параметр скорости обучения γ (число от 0 до 1).

2. На вход сети подается вектор $\mathbf{x}_n$ и определяется нейрон, веса которого наиболее подходят по значениям.

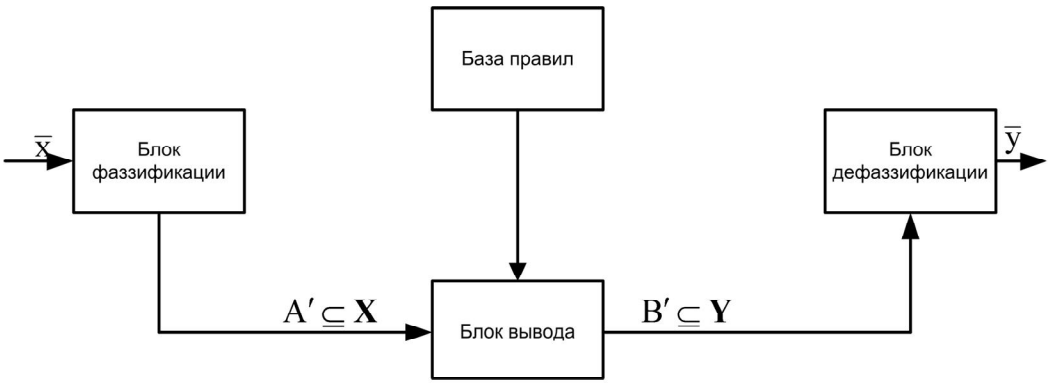

Рис. 3. Открытая модель модуля нечеткого управления

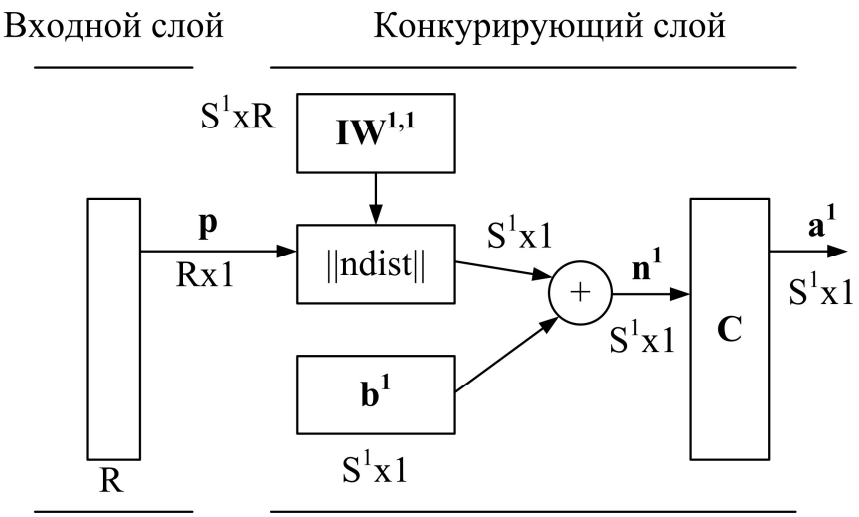

Рис. 4. Архитектура слоя Кохонена

3. Найденный нейрон становится "победителем", а вектор его весовых коэффициентов $\mathbf{w}_i$ вычисляется по формуле

$$\mathbf{w}_i(m+1) = \mathbf{w}_i(m) + \gamma\left(\mathbf{x}_n - \mathbf{w}_i(m)\right).$$

4. Все векторы предъявляются поочередно, вплоть до стабилизации нейросети.

С использованием слоев или карт Кохонена можно синтезировать нейронную сеть, позволяющую осуществить кластеризацию любой совокупности векторов.

В третьей главе рассматривается разработанное алгоритмическое обеспечение для автоматизированного категорирования персональных данных.

В качестве основы для внедрения механизмов автоматизированного категорирования ПДн в уже существующую систему автоматического проектирования баз данных будем использовать следующий базовый алгоритм (рис. 5).

Важной составляющей представленного алгоритма является собственно алгоритм автоматизированного категорирования ПДн. Рассмотрим его в следующем виде (рис. 6).

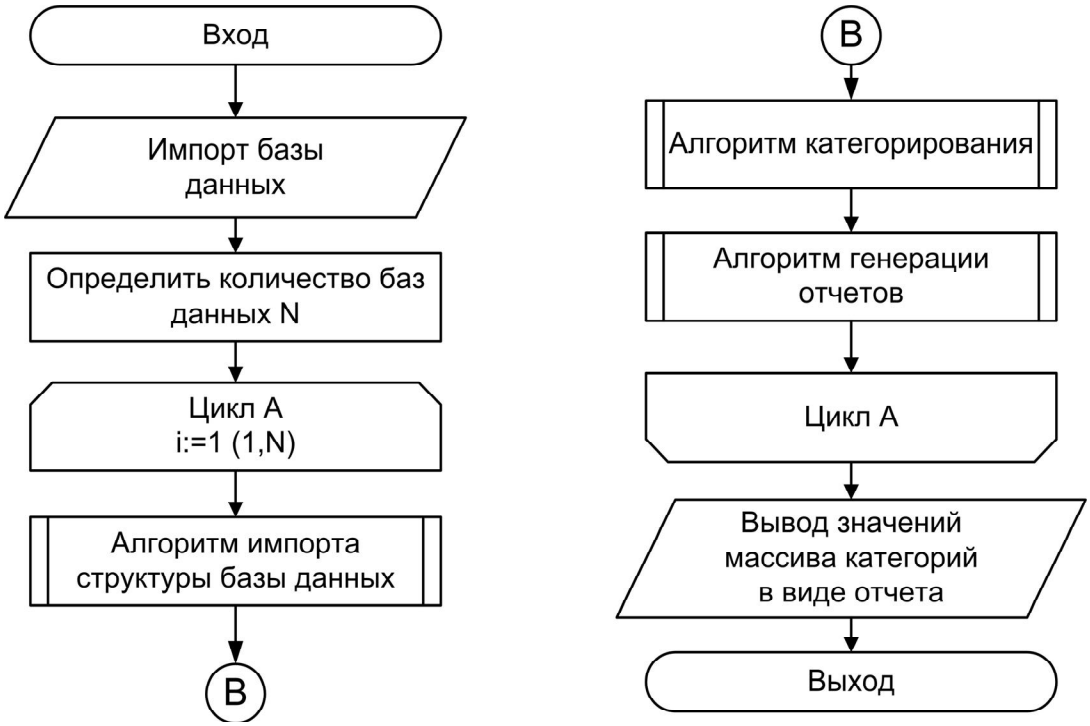

Рис. 5. Алгоритм внедрения механизмов автоматизированного категорирования ПДн

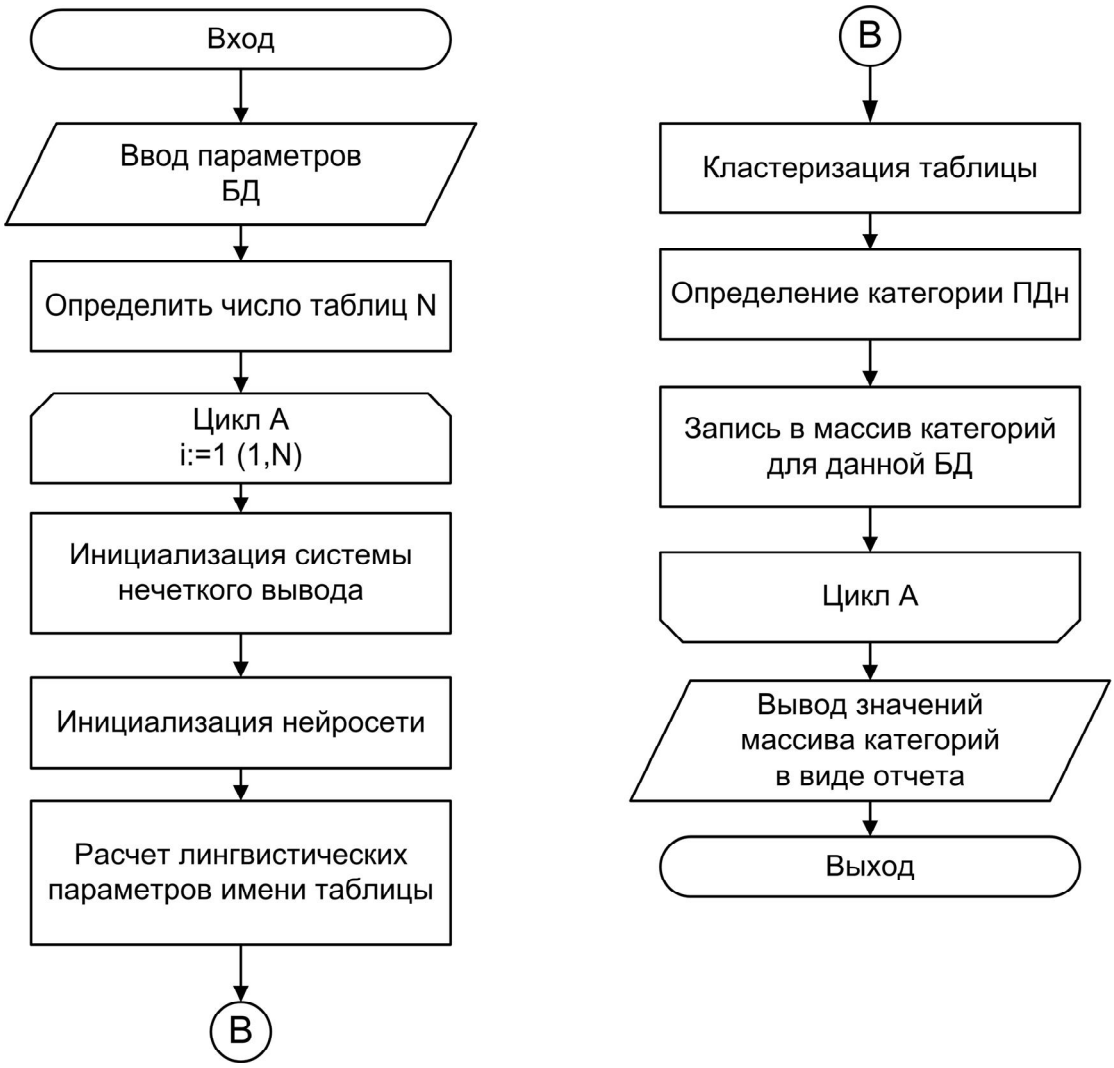

Рис. 6. Алгоритм автоматизированного категорирования ПДн

В третьей главе также рассмотрены основные UML-модели информационной системы (ИС) автоматизированного категорирования ПДн, предложен вариант автоматизации проектирования подобной ИС с помощью пакета Rational Rose Enterprise Edition фирмы IBM. В частности, следует выделить следующие наиболее важные моменты.

Статические модели обеспечивают представление структуры информационных систем, не отображая особенностей изменений системы во времени. В качестве статической модели может выступать диаграмма классов нейросетевой системы автоматизированного категорирования (рис. 7).

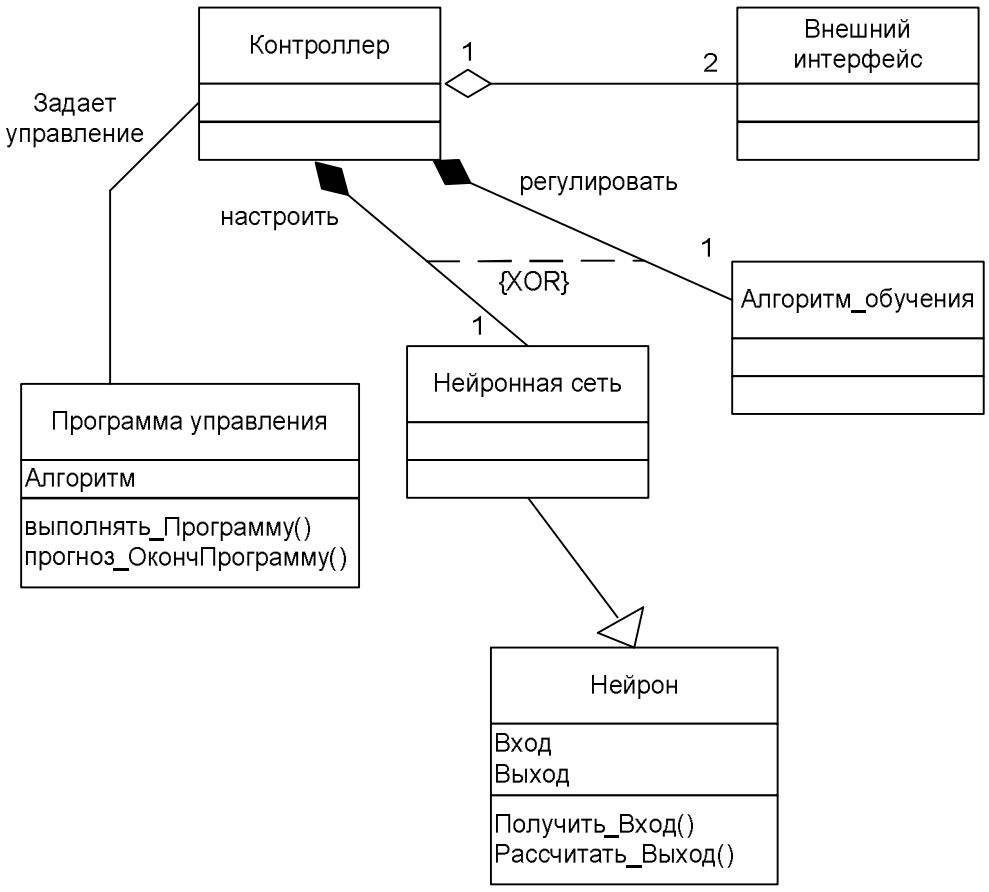

Рис. 7. Статическая модель нейросетевой системы автоматизированного категорирования

Динамические модели отражают изменение состояний в процессе работы информационной системы с течением времени. Для построения таких моделей, как правило, используют диаграммы Use Case и диаграммы последовательности. Рассмотрим фрагмент диаграммы последовательности для процесса простого нейросетевого анализа при категорировании ПДн (рис. 8).

Данные модели позволяют усовершенствовать традиционную систему принятия решений за счет автоматизации категорирования ИСПДн и посредством искусственного (нейросетевого) интеллекта сделать подобную систему более гибкой и оперативной. Вследствие этого повышается скорость модификации СЗИ для ИСПДн, уменьшаются временные интервалы "информационной неопределенности" и повышается суммарная оценка защищенности ПДн.

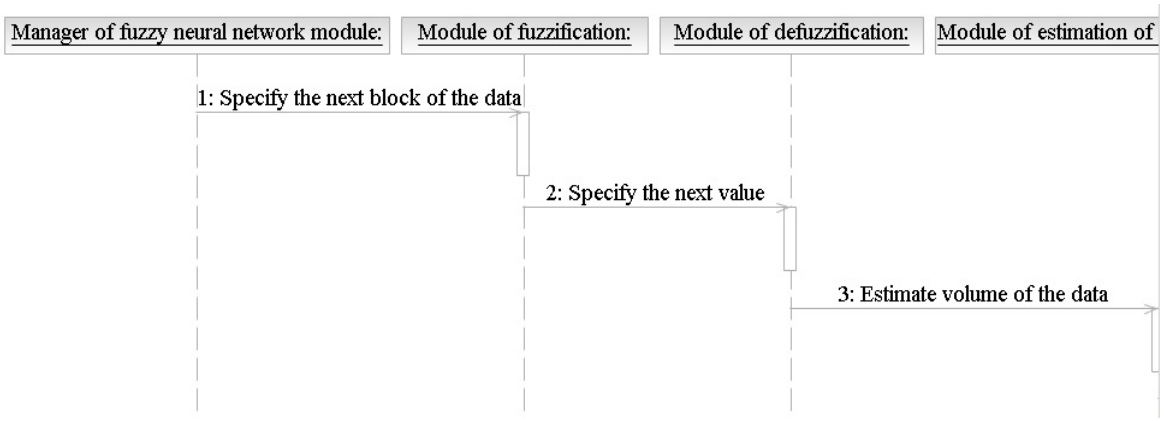

Рис. 8. Фрагмент диаграммы последовательности
для простого нейросетевого анализа

В четвертой главе предложены варианты программной реализации системы автоматического категорирования ПДн в рамках систем управления информационной безопасностью ПДн (СУИБ ИСПДн).

В частности, для системы дистанционного обучения (СДО) разработана система, реализованная на базе Web-платформы с применением технологий PHP и MySQL.

Анализ производительности функционирования СУИБ для ИСПДн типа СДО позволяет выявить следующие закономерности в изменениях параметров (K1, L1 – качество результата и длительность процедуры при простом анализе, K2, L2 – качество результата и длительность процедуры при глубоком анализе), представленные в таблице 1.

Таблица 1. Показатели производительности работы СУИБ для СДО

	K1	L1	K2	L2
Server 1	90%	0,128 с	100%	0,244 с
Server 2	90%	0,129 с	100%	0,247 с
Server 3	90%	0,127 с	100%	0,245 с
Server 4	90%	0,126 с	100%	0,246 с
Server 5	90%	0,128 с	100%	0,244 с

Для кадровой системы предприятия (КСП) разработана система управления, реализованная на базе сервера баз данных Interbase 6 с применением технологий объектно-ориентированного программирования в системе Borland Delphi 7.

Анализ производительности СУИБ позволяет выявить следующие закономерности в изменениях параметров (K1, L1 – при простом анализе, K2, L2 – при глубоком анализе), представленные в таблице 2.

Таблица 2. Показатели производительности работы СУИБ для КСП

	K1	L1	K2	L2
1	90%	3,321 с	100%	5,321 с
2	90%	3,425 с	100%	5,425 с
3	90%	3,369 с	100%	5,369 с
4	90%	3,350 с	100%	5,350 с
5	90%	3,324 с	100%	5,324 с

Следует заключить, что эффективность нейросетевого анализатора при втором виде анализа несколько выше, чем при первом виде. Однако производительность работы СУИБ ИСПДн снижается почти вдвое. Поэтому для относительно небольших баз данных рекомендуется использовать глубокий нейросетевой анализ, а для больших – сначала простой, а при повышении требований к информационной безопасности – глубокий нейросетевой анализ.

При глубоком анализе экспериментальным путем была выявлена общая оценка эффективности автоматизации категорирования персональных данных (по оси Ox – количество полей таблиц, по оси Oy – количество секунд).

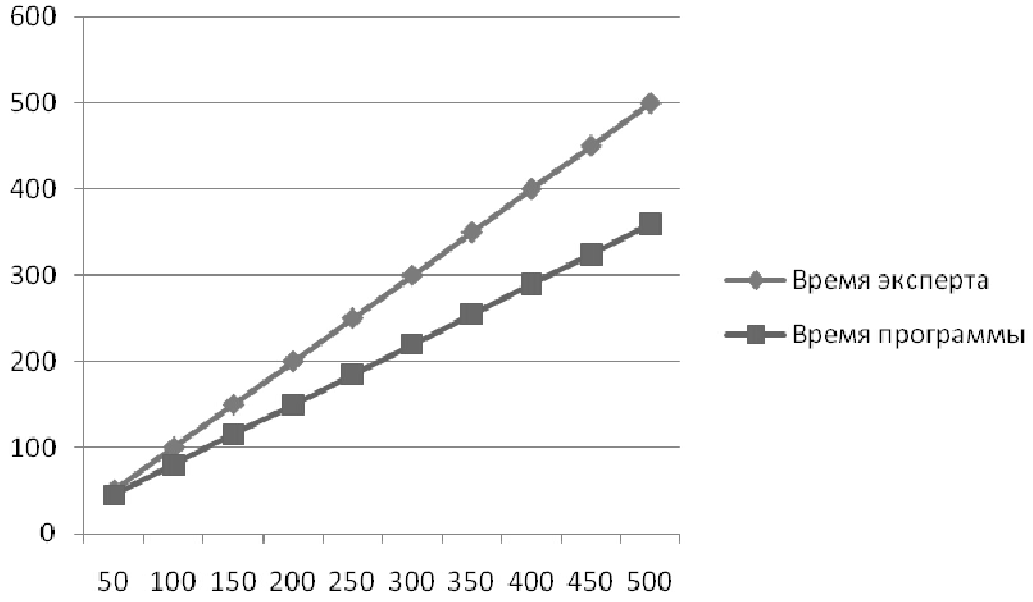

Рис. 9. График анализа эффективности автоматизации категорирования ПДн при глубоком анализе

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

В диссертационной работе решена актуальная научно-техническая задача автоматизированного категорирования персональных данных в информационных системах. Основными научными и практическими результатами диссертационной работы являются следующие:

1. Разработаны алгоритмы для автоматизированного категорирования персональных данных. В частности, разработаны алгоритмы для реализации процессов кластеризации блоков персональных данных, работы нейронной сети, получения числовых характеристик блоков персональных данных.

2. Разработаны специальные модификации стандартных нейронных сетей для создания систем автоматизированного категорирования ПДн на основе систем нечеткого вывода. Выявлены преимущества и недостатки существующих математических моделей нейронных сетей для категорирования данных. На основе исследованных моделей синтезирована новая модель нейронной сети, пригодной для высокоэффективной реализации категорирования персональных данных.

3. Разработаны специальные модификации систем нечеткого вывода Мамдани-Заде и Такаги-Сугено для создания систем автоматизированного категорирования ПДн.

4. Разработаны UML-модели систем нечеткого нейросетевого анализа баз данных ИСПДн.

5. Разработан проект системы автоматизированного категорирования ПДн на основе нечетких нейросетевых технологий. Определены основные функции и синтезирована объектно-ориентированная программная архитектура системы автоматизированного категорирования ПДн.

6. Разработаны программные модули системы автоматизированного категорирования ПДн. Программные модули системы автоматизированного категорирования ПДн синтезированы для двух вариантов ИСПДн: web-сервер и внутрикорпоративный SQL-сервер.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях, рекомендованных ВАК

1. Корнев П.А. Применение нейронной сети Кохонена для управления информационной безопасностью персональных данных / Корнев П.А., Ма-

лыш В.Н. // Системы управления и информационные технологии. – 2011. – №4(46). – С. 53-58.

Публикации в других изданиях

2. Корнев П.А. Основные принципы управления системой информационной безопасности организации / Зияутдинов В.С., Корнев П.А. // Совершенствование системы управления организацией в современных условиях: сборник материалов V Всероссийской научно-практической конференции. – Пенза, 2008. – С. 78-81.

3. Корнев П.А. Математические основы аппроксимационных свойств нейронных сетей / Зияутдинов В.С., Корнев П.А., Малыш В.Н. // Математические методы и информационные технологии в экономике, социологии и образовании: сборник материалов XXIV Международной научно-технической конференции. – Пенза, 2009. – С. 82-85.

4. Корнев П.А. Применение пакета MATLAB для создания, обучения и тестирования нейронных сетей / Зияутдинов В.С., Корнев П.А., Малыш В.Н. // Актуальные проблемы современного образования: сборник материалов межрегиональной научно-методической конференции. – Воронеж, 2009. – С. 184-187.

5. Корнев П.А. Основные стадии решения стандартных задач с использованием нейронных сетей / Зияутдинов В.С., Корнев П.А. // Инновации и информационные технологии в образовании: сборник научных трудов II Всероссийской научно-практической конференции: в 2ч. – Липецк: ЛГПУ, 2010. – Ч. 2. – С. 42-44.

6. Корнев П.А. UML-моделирование систем нейросетевого управления СУБД / Зияутдинов В.С., Корнев П.А. // Актуальные вопросы технических, экономических и гуманитарных наук: Материалы III международной заочной научно-практической конференции. – Георгиевск: Георгиевский технологический институт (филиал) ГОУ ВПО "Северо-Кавказский государственный технический университет", 2010. – С. 23-27.

7. Корнев П.А. Основы автоматизации кластеризации данных с применением нейросетей Кохонена / Корнев П.А. // Информационно-вычислительные технологии и их приложения: сборник статей XII Международной научно-технической конференции: МНИЦ ПГСХА. – Пенза: РИО ПГСХА, 2010. – С. 100-103.

8. Корнев П.А. Основы формализации информационных потоков в процессах защиты персональных данных / Зияутдинов В.С., Корнев П.А. // Информационные технологии в процессе подготовки современного специалиста. Межвузовский сборник научных трудов. – Выпуск 14. – Липецк, 2011. – С. 57-61.

9. Корнев П.А. Логико-лингвистическое описание систем управления информационной безопасностью персональных данных / Корнев П.А., Малыш В.Н. // Информационные технологии в процессе подготовки современного специалиста. Межвузовский сборник научных трудов. – Выпуск 14. – Липецк, 2011. – С. 80-83.

10. Корнев П.А. Нечеткое нейросетевое управление информационной безопасностью персональных данных / Корнев П.А. // Математическое и программное обеспечение вычислительных систем: Межвуз. сб. науч. тр. / Под ред. А.Н. Пылькина – Рязань (РГРТУ), 2011. – С. 72-79.

11. Корнев П.А. Программа «NNGuardPDw», номер государственной регистрации №2012610246 от 10.01.2012 г.

12. Корнев П.А. Программа «NNGuardPDI», номер государственной регистрации №2012611845 от 17.02.2012 г.

13. Корнев П.А. Нечеткие операторы как инструменты агрегирования нечетких персональных данных / Корнев П.А. // Перспективы развития информационных технологий: сборник материалов VI Международной научно-практической конференции. – Новосибирск: Издательство «СИБПРИНТ», 2012. – С. 130-135.