В большинстве информационных систем используется механизм идентификации и аутентификации на основе схемы идентификатор пользователя / пароль. Это самый простой, однако и самый ненадежный и уязвимый способ. Однако существуют и другие альтернативные и более защищенные типы механизмов защиты, которые могут быть реализованы для обеспечения службы идентификации и аутентификации:

– механизм, основанный на паролях – уязвимый и не обеспечивающий достаточную надежность;

– механизм, основанный на интеллектуальных картах – смарт-карта реализует аутентификацию с помощью схемы – запрос/ответ в реальном масштабе времени, что помогает предотвратить получение злоумышленником неавторизованного доступа путем воспроизведения сеанса регистрации пользователя;

– механизм, основанный на биометрии – базируется на опознавании пользователя по сугубо индивидуальным характеристикам;

– механизмы блокировки ПК или автоматизированного рабочего места – позволяет пользователям оставаться зарегистрированными, не делая при этом свое место потенциально доступным для злоумышленников;

– завершение соединения после нескольких ошибок при регистрации – позволяет исключить подбор пароля с помощью перебора возможных комбинаций;

– уведомление пользователя о "последней успешной регистрации" и "числе ошибок при регистрации" – информирует пользователя об использовании его регистрационного имени.

Методы, основанные на биометрии, достаточно сложны и требуют специального оборудования. Известны следующие методы:

– персональные: отпечатки пальцев, строение лица;

– квазистатические: геометрия руки, особенность глаз, отпечатки ладоней, рисунок кровеносных сосудов;

– квазидинамические: пульс, баллистокардиография, энцефалография;

– динамические: голос, почерк, стиль печатания.

Применение этих механизмов в разной мере необходимо для обеспечения предотвращения несанкционированного использования ресурсов системы и данных в зависимости от степени конфиденциальности защищаемой информации и принятой политики безопасности.

3. Выводы и перспективы дальнейших разработок

Научная новизна: результатом исследования является классификация и выделение существующих технологий современных средств защиты, их достоинств и недостатков, а также сферы применения. Это используется при комплексной оценке выбора средств защиты информации.

Литература: **1.** *Страссберг Кейт Е., Гондек Ричард Г., Роли Гари*. Полный справочник по брандмауэрам. М.: Вильямс, 2004. 836 с. **2.** *Домарев В.В*. Безопасность информационных технологий. Методология создания систем защиты. К.: ООО "ТИД "ДС", 2001. 688 с. **3.** *Филимонов А.* Протоколы Интернета. СПб.: изд. "BHV-СПб", 2003. 516 с. **4.** *Медведовский И.Д., Семьянов П.В., Платонов В.В*. Атака через Интернет. НПО "Мир и семья-95", 1997. **5.** *Козлов Д.А., Парандовский А.А., Парандовский А.К.* Энциклопедия компьютерных вирусов. М.: "СОЛОН-пресс", 2001. 464 с.

Поступила в редколлегию 19.01.2007

Рецензент: д-р техн. наук, проф. Самойленко Н.И.

Дударь Зоя Владимировна, канд. техн. наук, проф., зав. кафедрой программного обеспечения ЭВМ, декан ФПО ХНУРЭ. Научные интересы: математическое и программно-техническое обеспечение взаимодействия крупномасштабных систем баз данных в динамическом окружении, дистанционное образование. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, тел. 7021-446.

Збитнева Майя Вячеславовна, канд. техн. наук, старший преподаватель кафедры программного обеспечения ЭВМ ХНУРЭ. Научные интересы: автоматизированные системы диспетчерского управления электрическими сетями, стеганография, интеллектуальные агенты, программирование под Web. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, тел. 7021-446.

УДК681.30001.571

ИСПОЛЬЗОВАНИЕ ИСКУССТВЕННОЙ НЕЙРОННОЙ СЕТИ В ПОДСИСТЕМЕ ВВОДА ГОЛОСОВОЙ ИНФОРМАЦИИ САПР ТП РОБОТИЗИРОВАННОГО ПРОИЗВОДСТВА

НЕВЛЮДОВ И.Ш., ЦЫМБАЛ А.М., МИЛЮТИНА С.С.

Для голосового задания управляющих команд робота предлагается использование искусственной нейронной сети (ИНС). Разрабатывается программное обеспечение, реализующее многослойный персептрон на основе переходных функций различного вида. Полученные результаты используются при создании подсистемы ввода голосовой информации САПР технологических процессов роботизированного производства.

1. Введение

Техническое зрение и тактильное очувствление призваны повысить возможности робота в восприятии информации об изменениях внешней среды. Применение средств адаптации роботов позволяет освободить человека от выполнения однообразных, повторяющихся операций и возложить на него исполнение более сложных в интеллектуальном отношении задач, например, по наблюдению за работой роботов, выполняющих эти операции.

Командовать машинами человек может не только с помощью кнопок, клавиатур, тумблеров управления, передающих цифровые или аналоговые сигналы кон-

троллерам роботов. Много преимуществ можно получить, если для передачи команд использовать человеческий голос, при этом:

– снижается усталость работающего;

– повышается скорость и гибкость передачи команд;

– высвобождаются руки для выполнения других функций (например, для заметок о ходе выполнения технологического процесса);

– передается более насыщенная, богатая по содержанию информация в ответ на возникшую ситуацию;

– уменьшается однообразие работы, поскольку оператор может использовать свой орган слуха для контроля правильности подаваемых команд, тем самым более активно вовлекаясь в рабочий процесс [1];

– становится возможным осуществление бесконтактного управления различными системами;

– появляется возможность управления сложными комплексами в опасных для человека условиях [2].

Подсистема ввода голосовой информации может быть частью САПР технологических процессов роботизированного производства. Таким образом, проведение исследований в области применения голосового управления роботами позволит достичь улучшения качества проектирования и поддержки технологических процессов, что указывает на актуальность и своевременность данного исследования.

2. Постановка задачи

Целью работы является описание основных особенностей разработки системы анализа голосовой информации в составе САПР технологических процессов роботизированного производства с использованием нейронной сети, что позволит сократить время и стоимость проектирования на 15-20%.

Поставленная цель достигается решением следующих задач: разработка программной модели многослойного персептрона, анализ процесса обучения голосовым командам при различной топологии нейронной сети, отработка распознавания управляющих голосовых команд, анализ и формирование технологических операций и переходов промышленного робота. Разработка программного обеспечения производится в среде программирования Visual C++.

3. Основные принципы создания системы анализа голосовой информации

Распознавание речи открывает широкие возможности применения его в робототехнике. Одна из возможных реализаций системы распознавания речи представлена на рис.1. Оператор распознает сложную деталь (например, кремниевую пластину) на экране и командует роботу: «Загрузить палету А». Эта команда влечет за собой множество действий. Прежде всего, после соответствующей обработки звукового сигнала производится его кодирование и анализ его структуры. Как только произнесенное выражение обработано в виде лингвистической строки, начинается процесс распознавания смысла (понимания) речи [1].

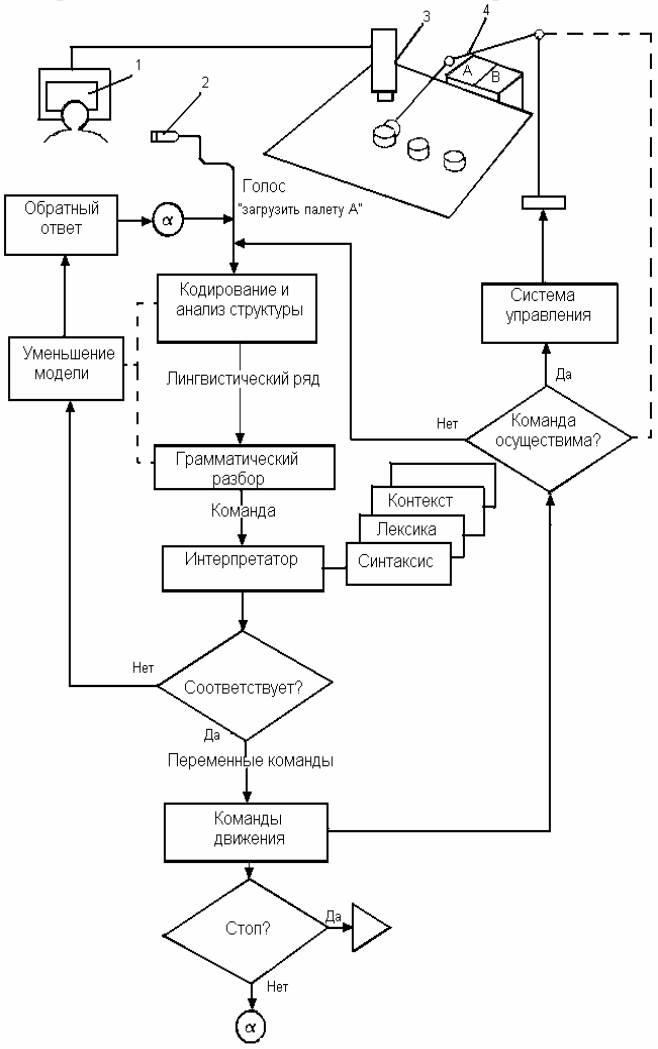

Рис. 1. Распознавание речи и система обработки естественного языка: 1 – монитор; 2 – микрофон; 3 – камера; 4 – робот [1]

При обработке естественного языка используется процедура разбивки фразы на распознаваемые ключевые слова путем грамматического разбора. Полученная структура команды затем исследуется по синтаксису (грамматике, времени), словарю и контексту. Если итоговый результат несовместим с известными машине «правилами», то можно уменьшить размерность модели и либо повторить весь процесс, либо снова сделать грамматический разбор строки. Если же команда введена правильно, т.е. она распознана и её возможно выполнить, то генерируется последовательность команд движения, которые являются входом в контроллер робота [1].

Если подаваемые команды несовместимы с текущим положением робота (например, робот не находится в положении, подходящем для взятия нужного объекта), компьютер вырабатывает звуковой сигнал обратной связи человеку (синтезирует речь), предлагая оператору выполнить корректирующие действия. Этот обратный процесс включает преобразование цифровой информации в точный аналоговый сигнал, управляющий громкоговорителем. Каков бы ни был ре-

зультат, речевая обратная связь необходима для того, чтобы оператор смог подтвердить (голосом) команду, перед тем как фактическое движение робота будет совершено [1].

В общем случае легче создать анализаторы речи, распознающие говорящего или реагирующие на отдельные слова, чем машины, распознающие связную речь. Большие трудности для распознавания речи создают голосовые изменения, присущие речи разных людей или одного человека. Структура предложений, значения фраз и морфология речи обязательно должны быть запрограммированы, в частности, в виде «правил» для машины с ИИ, чтобы позволить роботу самообучаться в процессе ведения разговора [1].

В современных компьютерных системах все больше внимания уделяют построению интерфейса с естественным вводом-выводом информации (распознавание рукописного текста, речевой диалог).

Наиболее перспективными на сегодняшний день являются системы речевого ввода. Задачу распознавания речевой информации можно разделить на две большие подзадачи:

– непосредственное распознавание отдельных слов;

– распознание смысла команд.

Непосредственное распознавание отдельных слов осложняется рядом факторов: различием языков, спецификой произношения, шумами, акцентами, ударениями и т.п.

В настоящее время можно выделить два основных направления при построении систем распознавания речи.

Эталонный метод основан на сравнении некоторых характеристик речи (энергетических, спектральных и т.п.). В качестве эталонов в большинстве случаев используют целые слова. Данный метод удобен для применения в системах с ограниченным словарем (например, для ввода небольшого набора команд) [2]. Эталоны формируются путем статистической обработки большого числа шаблонов. Сравнение входного сигнала с эталоном возможно путем нечёткого сопоставления образов [3].

Фонемно-ориентированный метод основан на выделении фонем из потока речи.

Сравнивая распознавание речевого потока методом распознавания целых слов и распознавание фонем, можно сделать вывод: при небольшом количестве слов, используемых оператором, более высокую надежность и скорость можно ожидать от распознавания целых слов, но при увеличении словаря скорость резко падает. Предположительно, размер словаря системы распознавания уже в сотню слов делает переход на уровень более низкий, чем распознавание слов в целом, актуальным [3].

Естественно, что перед тем, как робот получит команду на выполнение, должны быть тщательно проверены ее правильность и осуществимость (либо обеспечена мгновенная реакция на звуковую команду прекратить движение). Наиболее эффективной проверкой правильности выполнения роботом требуемой задачи было бы графическое представление в режиме «офф-лайн» (в автономном режиме) [1].

4. Применение методов искусственных нейронных сетей в системах распознавания речи

Одним из видов нейронных сетей являются обучаемые сети. Этот вид сетей используют для неформализуемых задач, к разряду которых относится распознавание речи. В процессе обучения сети автоматически изменяются такие её параметры, как коэффициенты синаптических связей, а в некоторых случаях и топология [4].

Нейронные сети также имеют свойство классификации объектов по их числовым параметрам. При обучении сети с учителем можно научить ее распознавать объекты, принадлежащие заранее определенному набору классов. Если же сеть обучается без учителя, то она может группировать объекты по классам в соответствии с их цифровыми параметрами.

Таким образом, на базе нейронных сетей можно создавать обучаемые и самообучающиеся системы.

Использование нейронной сети, как одного из средств реализации интеллектуального анализа данных, позволяет решить следующие задачи:

– моделировать сложные нелинейные зависимости между данными и целевыми показателями;

– выявлять тенденции в данных (при наличии временных рядов) для построения прогнозов;

– работать с зашумленными и неполными данными;

– получать содержательные результаты при относительно небольшом объёме исходной информации с возможностью усовершенствования модели по мере поступления новых данных;

– выявлять аномальные данные, значительно отклоняющиеся от «открытых» устойчивых закономерностей, и т.д. [3].

Возможность создания на базе искусственных нейронных сетей самообучающихся систем является важной предпосылкой для их применения в системах распознавания речи.

После выделения информативных признаков речевого сигнала они представляются в виде некоторого набора числовых параметров. Далее задача распознавания примитивов речи (фонем и аллофонов) сводится к их классификации при помощи обучаемой нейронной сети [4].

Нейронные сети можно использовать и в более высоких уровнях распознавания слитной речи для выделения слогов, морфем и слов, что, как уже было отмечено, является более целесообразным при обучении ограниченному количеству слов-команд.

Для реализации системы голосового управления роботом предлагается использовать многослойный персептрон со скрытым слоем, который описывается функцией гиперболического тангенса в первом случае и сигмоидной функцией – во втором. Сигмоидная функция имеет следующий вид:

$$y = 1/(1+\exp(-(\sum_i W_i x_i - \Theta))), \quad (1)$$

где w_i – веса (коэффициенты синапса), на которые умножаются входные значения; x_i – входные значения для данного слоя; Θ – некий входной порог.

Ниже приведена функция гиперболический тангенс:

$$y = \tanh(x). \quad (2)$$

Для коррекции ошибок используется алгоритм обратного распространения. В соответствии с этим алгоритмом [3]:

1. Начальные значения весов всех нейронов всех слоев V(t=0) и W(t=0) полагаются случайными числами.

2. Сети предъявляется входной образ X^α, в результате формируется выходной образ $y \neq Y^\alpha$. При этом нейроны последовательно от слоя к слою функционируют по следующим формулам:

– скрытый слой:

$$x_i = \sum_i W_{ij} X_i^\alpha, \quad (3)$$

$$y_i = f(x_i), \quad (4)$$

– выходной слой:

$$x_k = \sum_j V_{jk} y_j, \quad (5)$$

$$y_k = f(x_k), \quad (6)$$

где f(x) – переходная функция слоя.

3. Функционал квадратичной ошибки сети для данного входного образа имеет вид:

$$E = 1/2(\sum_k (y_k - Y_k^\alpha)^2), \quad (7)$$

здесь y_k – реальный выход сети; Y_k^α – желаемое значение.

Данный функционал подлежит минимизации. Классический градиентный метод оптимизации состоит в итерационном уточнении аргумента согласно формуле:

$$V_{jk}(t+1) = V_{jk}(t) - h * \frac{\partial E}{\partial V_{jk}}, \quad (8)$$

где h – скорость обучения.

Функция ошибки в явном виде не содержит зависимости от веса V_{jk}, поэтому воспользуемся формулами неявного дифференцирования сложной функции:

$$\frac{\partial E}{\partial y_k} = \delta_k = y_k - Y_k^\alpha; \quad (9)$$

$$\frac{\partial E}{\partial x_k} = \frac{\partial E}{\partial y_k} * \frac{\partial y_k}{\partial x_k} = \delta_k * y_k(1-y_k), \quad (10)$$

$$\frac{\partial E}{\partial V_{jk}} = \frac{\partial E}{\partial y_k} * \frac{\partial y_k}{\partial x_k} * \frac{\partial x_k}{\partial V_{jk}} = \delta_k * y_k(1-y_k) * y_j. \quad (11)$$

Здесь учтено полезное свойство сигмоидной функции f(x): ее производная выражается только через само значение функции, f'(x)=f(1-f). Таким образом, все необходимые величины для подстройки весов выходного слоя V получены.

4. На этом шаге выполняется подстройка весов скрытого слоя. Градиентный метод по-прежнему дает:

$$W_{ij}(t+1) = W_{ij}(t) - h * \frac{\partial E}{\partial W_{ij}}. \quad (12)$$

Вычисления производных выполняются по тем же формулам, за исключением некоторого усложнения формулы для ошибки δ_j:

$$\frac{\partial E}{\partial y_j} = \delta_j = \sum_k (\frac{\partial E}{\partial x_k} * \frac{\partial x_k}{\partial y_j}) = \sum_k \delta_k * y_k(1-y_k) * V_{jk}, \quad (13)$$

$$\frac{\partial E}{\partial W_{ij}} = \frac{\partial E}{\partial y_j} * \frac{\partial y_j}{\partial x_j} * \frac{\partial x_j}{\partial W_{ij}} = \delta_j * y_j(1-y_j) * X_i^\alpha = \sum_k (\delta_k * y_k(1-y_k) * V_{jk}) * (y_j(1-y_j) * X_i^\alpha). \quad (14)$$

При вычислении δ_j был применен принцип обратного распространения ошибки: частные производные берутся только по переменным последующего слоя. По полученным формулам модифицируются веса нейронов скрытого слоя. Если в нейронной сети имеется несколько скрытых слоев, процедура обратного распространения применяется последовательно для каждого из них, начиная со слоя, предшествующего выходному, и далее до слоя, следующего за входным. При этом формулы сохраняют свой вид с заменой элементов выходного слоя на элементы соответствующего скрытого слоя.

5. Шаги 2-4 повторяются для всех обучающих векторов. Обучение завершается по достижении малой полной ошибки или максимально допустимого числа итераций [5].

5. Экспериментальные исследования

В ходе проведенных исследований разработано программное обеспечение, моделирующее работу многослойного персептрона. В качестве переходной функции скрытого слоя выбрана сигмоидная функция. Первоначально на вход сети подаётся сигнал, соответствующий некоторому слову – команде управления роботом.

Необходимо отметить, что при цифровой записи звуковой информации используются следующие частоты дискретизации сигнала: 11,025 кГц, 22,05 кГц, 44,1 кГц. Наличие высокой частоты значительно усложняет работу программного обеспечения, поэтому для первого эксперимента была использована частота дискретизации, равная 11,025 кГц.

Принятые сигналы управления пропускаются через фильтр и только после этого подаются на вход нейросети. Входной сигнал, представленный в векторном виде, взвешивается с помощью коэффициентов первого синапса (представлены в виде матрицы): коэффициенты синапса умножаются на значения сигнала. Результатом операции также является вектор. Далее сигнал подвергается обработке в скрытом слое, т.е. к полученному взвешенному сигналу применяется сигмоидная функция (результат – также вектор). Затем полученный вектор взвешивается с помощью коэффициентов второго синапса (строка): коэффициенты второго синапса умножаются на вектор. Результатом является одно число, к которому на завершающем этапе применятся сигмоидная функция. Полученный результат сравнивается с желаемым значением. Разница составляет ошибку обучения. Если она превышает какое-то допустимое значение, корректируются коэффициенты первого и второго синапсов, и исходный сигнал опять пропускается через нейросеть. Процесс продолжается либо до тех пор, пока ошибка не станет допустимой (появляется сообщение, что сеть обучена), либо пока не пройдёт максимальное количество этапов обучения.

Ниже приведены примеры сигналов, которым обучается нейросеть, и графики обучения (рис.2).

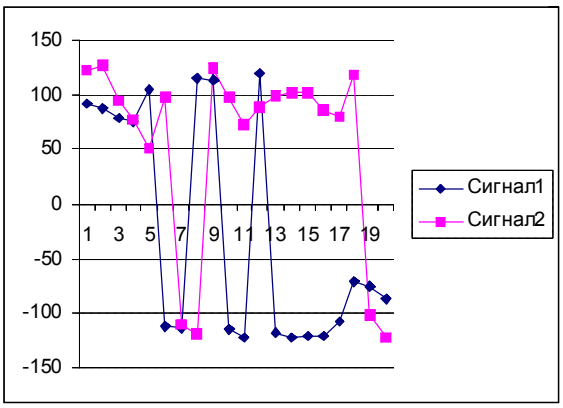

Рис. 2. Графики фрагментов сигналов

Сигнал 1 соответствует фрагменту слова «вперёд», сигнал 2 – фрагменту слова «стоп». Желаемое значение на выходе нейросети для сигнала 1 было установлено равным 0, для сигнала 2 – +0,5. Величина допустимого отклонения была выбрана равной 0,2. Таким образом, если на выходе нейросети при обучении получается значение меньшее или равное 0,2, то считается, что сеть обучена сигналу1. Аналогично для сигнала 2: если на выходе значение попадает в диапазон [0.3, 0.7], считается, что сеть обучилась второму сигналу.

Графики обучения фрагментам сигналов приведены на рис. 3. По оси абсцисс отложен номер шага обучения, по оси ординат – величина ошибки. На графике надпись Сигнал 1 обозначает значение ошибки для сигнала 1, Сигнал 2 – ошибку для сигнала 2.

Из графика видно, что при обучении сигналу 1 допустимое значение ошибки достигается на шестом шаге, а при обучении сигналу 2 – на девятнадцатом.

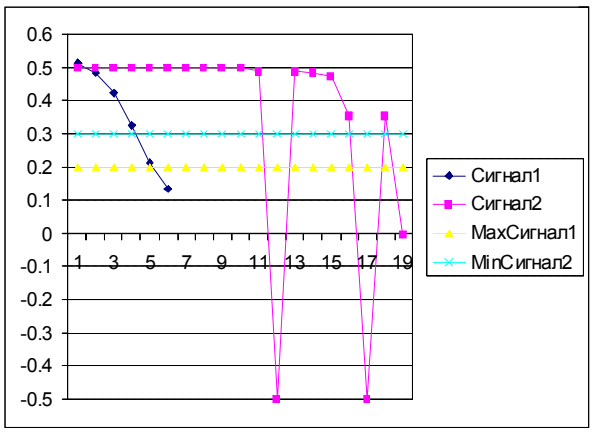

Рис. 3. Графики обучения фрагментам сигналов

6. Выводы

На сегодняшний день существует три метода программирования роботов: в режиме обучения, на языке программирования роботов и аналитическое программирование.

Применение голосового задания команд управления роботом позволит значительно упростить процесс управления и его программирование.

В ходе проведенных исследований разработано программное обеспечение для анализа голосовой информации с использованием искусственной нейронной сети. В качестве базовой модели нейронной сети был выбран многослойный персептрон. В целях интеграции модели в систему управления роботом персептрон реализован методами языка С++.

Результаты тестирования показали принципиальную возможность обучения нейронной сети отдельным словам-командам и их распознавания. Полученные результаты и разработанная программа могут использоваться для создания подсистемы ввода голосовой информации САПР технологических процессов роботизированного производства.

Научная ценность исследования состоит в разработке алгоритмического и программного обеспечения,

реализующего модель многослойного персептрона, используемого для распознавания голосовых команд. Применение голосового управления позволит сократить время и стоимость проектирования на 15-20%.

Практическая ценность работы заключается в уменьшении времени и стоимости технологической подготовки роботизированного производства за счёт голосового задания управляющих сигналов.

В дальнейшем планируется интегрировать разработанное программное обеспечение в систему управления роботами MR-999e и PM-01, также реализовать голосовое управление при помощи других моделей представления ИНС.

Литература: 1. *Искусственный* интеллект: Применение в интегрированных производственных системах / Под ред. Э. Кьюсиака: Пер. с англ. А.П.Фомина /Под ред. А.И. Дащенко, Е.В. Левнера. М.: Машиностроение, 1991. 544с. **2.** *Рабинер Л., Гоулд Б.* Теория и применение цифровой обработки сигналов. М.: Мир, 1978. **3.** *Киедзи Асаи, Дзюндзо Ватада, Сокуке Иваи* и др. Распознавание речи // Прикладные нечёткие системы /Под ред. Тэрано Т., Асаи К., Сугено М.. М.: Мир, 1993. **4.** *Комарцова Л.Г., Максимов А.В.* Нейрокомпьютеры. М.:Издательство МГТУ им. Н.Э.Баумана, 2002. 320с. **5.** *Терехов С.А.* Лекции по теории и приложениям нейронных сетей. 1994. Лаборатория Искусственных Нейронных Сетей НТО-2, ВНИИТФ, Снежинск **6.** *Нейроинформатика* / А. Н. Горбань, В. Л. Дунин-Барковский, А. Н. Кирдин, Е. М. Миркес, А. Ю. Новоходько, Д. А. Россиев, С. А. Терехов и др. Новосибирск: Наука, 1998. 296 С. **7.** *Головко В. А.* Нейронные сети: обучение, организация и применение. М.: ИПРЖР, 2001.

Поступила в редколлегию 14.02.2007

Рецензент: д-р техн. наук, проф. Ильченко Б.С.

Невлюдов Игорь Шакирович, д-р техн. наук, проф. ХНУРЭ. Научные интересы: технология приборостроения, гибкие производственные системы, робототехника. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, тел. (057)702-14-86.

Цымбал Александр Михайлович, канд. техн. наук, доцент, докторант ХНУРЭ. Научные интересы: системы программирования, системы искусственного интеллекта. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, тел. (057)702-14-86, e-mail: mcdulcimer@kture.kharkov.ua.

Милютина Светлана Святославовна, аспирантка кафедры ТАПР ХНУРЭ. Научные интересы: системы программирования, системы искусственного интеллекта. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, тел. (057)702-14-86.

УДК519.7

ОБЪЕКТНОЕ ПРЕДСТАВЛЕНИЕ ЭЛЕКТРОННЫХ ТЕКСТОВЫХ ДОКУМЕНТОВ

ГВОЗДИНСКИЙ А.Н., ГУБИН В.А.

Рассматривается проблема формализации содержимого электронных текстовых документов. Документы представляются в виде совокупности объектов двух видов – объектов контейнеров и атомарных объектов. Каждая группа объектов отражает либо топологический, либо информационный аспект содержимого документа.

Актуальность исследования

Бурное развитие вычислительной техники, сети Internet, приход компьютеров практически в каждый офис, в каждый дом порождает тенденцию увеличения удельного веса представления информации в электронном виде. С развитием концепции электронного документооборота на первый план выходят электронные документы как носители и источники информации, а документы на бумаге отходят на второй план, уступая свои позиции особенно в тех областях, где требуется высокий уровень мобильности и оперативности.

С другой стороны, бурное развитие сети Internet и ее общедоступность сделали практически неограниченным доступный информационный массив. Большая часть этого массива изначально не предполагала возможность автоматизированной обработки. Это породило необходимость перехода от методов обработки документов на бумажных носителях к развитию и совершенствованию технологий автоматизированной обработки электронных источников информации.

Данные обстоятельства привели к возникновению и развитию технологии Text Mining – современного направления интеллектуального анализа и обработки текстовых данных. Эта технология, являясь одним из направлений Data Mining, позволяет решать разнообразные задачи, возникающие при анализе больших электронных массивов неструктурированной информации.

Отличительной особенностью современных подходов в Text Mining является то, что единицей анализа содержимого электронных текстовых документов есть слово. При этом игнорируется то обстоятельство, что документы определенного класса могут состоять из текстовых фрагментов, обособленных относительно других фрагментов и представляющих ценность как некоторая неделимая единица. Для определенного класса задач, в частности, для задач идентификации данных в текстовых документах, это может быть достаточно существенным недостатком. Настоящая работа предлагает подход, устраняющий этот недостаток.

Целью исследования является формализация содержимого электронных текстовых документов [1]. При этом документы представляются в виде совокупности объектов двух видов – объектов контейнеров и атомарных объектов. Первая группа объектов отражает топологию документа, вторая – его информационное содержимое. Также важно, чтобы о каждом обособленном текстовом фрагменте документа сохранялась информация о контексте его появления.

Задачи исследования: разработка спецификации объектов контейнеров и атомарных объектов; разработка методики определения того, какие фрагменты исходного документа необходимо отнести к объектам того