Насыпный В.В.

СТОХАСТИКА

Перспективная информационная технология

Москва 2012

ББК 22
Н 34

Насыпный В.В. Стохастика. Перспективная информационная технология. - М.:МПГУ, 2012. - 106 с.

Книга посвящена стохастической информационной технологии или стохастики – одному из наиболее перспективных направлений развития компьютерных систем в 21 веке. Стохастика успешно отвечает на все вызовы в информационной технологии, включая решение проблемы BigData, создания индустрии знаний и квантового интеллекта, реферирования неструктурированных текстов, методов распознавания и понимания смысла речевой и видео- информации, гарантированной защиты вычислительных процессов в компьютере и др. Разработанные методы имеют мировой уровень новизны.

Книга предназначена для специалистов в области информатики, искусственного интеллекта, лингвистики и семиотики, квантовых компьютеров, защиты информации, а также для всех читателей, которых интересует перспективы развития компьютеров.

ISBN 978-5-94845-245-6

Введение

Научное открытие – стохастическая саморазвивающаяся система

Представленное научное издание посвящено одному из наиболее перспективных направлений развития компьютерных систем, которым является стохастическая информационная технология или стохастика. Стохастика была разработана в 90-х годах во время первого этапа создания систем искусственного интеллекта. Выяснилось, что традиционная технология не позволяет эффективно обрабатывать в среде современных компьютеров (в основе которых машина Тьюринга) сложные символьные конструкции. В результате не был реализован логический вывод - основа интеллектуальных систем. Этому мешал "комбинаторный взрыв", вызванный переборным способом его реализации. Появилась новая идея: свести сложные символьные конструкции с использованием стохастического (случайного) преобразования к случайным числам− комбинациям заданной (определенной) длины, обеспечивающей заданную сколь угодно малую вероятность коллизий при последующей обработке как элементарных символьных конструкций, так и их сочетаний. Эти числа назывались стохастическими индексами. По сути, они являлись новыми уникальными именами символьных конструкций любой сложности. Здесь проявилось первое свойство стохастики: случайность − уникальность. В итоге все конструкции получают уникальные случайные имена. Далее работает тезис: назвать значит понять.

Таким образом, в случайные индексы было введено знание исходных символьных конструкций (в закодированном виде). И здесь открылось совершенно новое качество стохастических систем − способность к саморазвитию и самообучению. Эти системы могли самостоятельно, без внешних воздействий производить путем сложения случайных индексов реализацию логических и семантических связей, таких как "часть-целое", "род-вид", "причина-следствие", "условие-заключение" и др. Например, при сложении по mod.2 стохастических индексов слов образовывался уникальный (со сколь угодно малой заданной вероятностью коллизий) случайный индекс словосочетания. При сложении случайных индексов словосочетаний формировался индекс предиката. При сложении по mod.2 индексов предикатов можно сформировать уникальный индекс предложения или правила продукций и т.д.

На основе этого впервые была получена саморазвивающаяся интеллектуальная среда, которая могла автоматически саморазвиваться, формируя при этом новые знания. Отсюда следует второе, очень мощное свойство стохастики − способность к самообучению путем автоматического формирования метазнаний, определяющих логическую связность семантически близких стохастических структур. В итоге автоматически реализуется неперебoрный метод логического вывода (независимо от объема исходного пространства поиска), использующий только логически связанные элементы знаний. Это позволило решить проблему комбинаторного взрыва и создать принципиально новое поколение саморазвивающихся, самообучающихся эффективных интеллектуальных систем, аналогов которым в мире сейчас не существует.

Как было отмечено выше, стохастика или стохастическая информационная технология обладает следующими свойствами, без которых невозможно понимание смысла и извлечение знаний из неструктурированных текстов:

- реализация логического вывода на больших пространствах поиска с использованием только логически и семантически связанных текстовых структур с исключением перебора на всем пространстве поиска, что обеспечивает исключение комбинаторного взрыва;

- осуществление саморазвития и самообучения системы новым знаниям, определяющим логически и семантически связанные элементы текста, формирование новых знаний, необходимых для получения «картины мира» и семантических классификаторов в различных проблемных областях;

- решение проблемы BigData путем автоматического создания баз знаний, описывающих все возможные свойства понятий и логических связей картины мира во всех возможных ситуациях на пространствах поиска объемом не менее 10^{15};

- реализация аналитических и поисковых функций на множестве исходной текстовой информации с использованием полученных знаний в реальном масштабе времени с максимальным пространством поиска не менее 10^{20}.

Указанными свойствами обладает только стохастика. Они не доступны для традиционных информационных технологий. Стохастика была разработана в России В.Насыпным и впервые опубликована в монографии «Развитие теории построения открытых систем на основе

информационной технологии искусственного интеллекта» (М.: Воениздат, 1994. - 248с.).

Кроме указанных свойств стохастика обладает еще двумя свойствами, которые являются актуальными для современного этапа развития IT, а именно:

- обеспечение гарантированной защиты создаваемых интеллектуальных систем, работающих в режиме BigData, от вредоносных информационных воздействий на основе возможности реализации всех поисковых и аналитических функций в зашифрованном виде на множестве зашифрованных текстовых массивов;

- полная совместимость стохастики с квантовой технологией, что обеспечивает возможность построения уже в настоящее время квантовых компьютеров с интеллектом, реализующих аналитические и поисковые функции на множестве пространства поиска не ниже 10^{30}.

Данная книга содержит специально подобранный комплекс материалов, сформированных на основе авторских публикаций разных лет и объединенных единым замыслом. Он предназначен для раскрытия содержания описанных выше уникальных возможностей стохастики.

1. Развитие и внедрение стохастики

В настоящее время в качестве необходимого условия развития информационной технологии рассматривается решение проблемы BigData. Теория и практика данной проблемы рельефно представлена в работе [1]. Несколькими годами раньше в России для решения проблемы обработки больших объемов данных и знаний в процессе развития искусственного интеллекта была создана стохастика – стохастическая информационная технология. Данный сборник содержит достаточно полное описание наиболее важных разделов стохастики, которые обосновывают возможность создания на ее основе нового перспективного направления информационных систем. Впервые стохастическая интеллектуальная информационная технология была представлена в работе автора этого сборника [4]. Указанная технология разрабатывалась в период с 1990 по 2000 годы [2-6].

Затем под научным руководством автора данного сборника на основе стохастики коллективом ученых были созданы проекты интеллектуальной поисковой системы [5], самообучающейся аналитической системы [6], комплексной защиты информации в компьютерах [2, 7, 11], распознавания и понимания слитной речи от неизвестного диктора [3, 8] и другие. Все они запатентованы в России и за рубежом.

Для реализации проекта интеллектуальной поисковой системы были организованы научно-производственные фирмы «Стокона» в России и «Stochasto» в Норвегии. Финансирование этого проекта осуществлялось международными инвестиционными фондами. В результате на основе российской стохастической информационной технологии была разработана и установлена в США (2005 год) первая в мире интеллектуальная поисковая система NearU. Эта система обеспечивала реализацию базовых функций лингвистической и семантической обработки больших объемов неструктурированной текстовой информации, данных и знаний. Система должна была осуществлять реализацию индуктивного логического вывода на больших объемах данных и знаний в реальном масштабе времени, а также извлечение знаний из текстов с использованием алгоритмов самообучения [5].

Таким образом, задолго до понимания проблемы больших данных (BigData) на Западе, в том числе лидерами IT технологий, например, Microsoft и Google, стохастика решила эту проблему.

Был также создан опытный образец на русском и английском языках для реализации в России в рамках создания интеллектуальной поисковой системы для Интернет - http://asknet.ru/

В настоящее время к разработке готовы следующие проекты: система распознавания и понимания смысла речи, интеллектуальная поисковая система, самообучающаяся аналитическая система, проактивная защита информации в компьютерах и др.

Отметим, что все эти проекты в той или иной степени решают проблему BigData, включая такие задачи, как накопление, ведение и логическая обработка больших объемов данных и знаний, извлечение знаний из речи, видеообразов и неструктурированной текстовой информации. Реализуются функции самообучения системы [6], глубокая аналитическая обработка неструктурированных текстов, данных и знаний.

Эти проекты являются уникальными и не имеют аналогов за рубежом. Внедрение их в России внесло бы существенный вклад в модернизацию экономики и промышленности в целом.

Из сказанного следует, что проблема BigData в России была поднята и решена на основе стохастики более, чем за десять лет до осознания ее значимости ведущими западными IT компаниями. Самое главное заключается в том, что именно в России впервые в мире исследовали и реализовали на основе стохастики проблему семантического анализа и понимания смысла сенсорной и текстовой информации как основу для извлечения знаний и глубокой аналитической обработки текстов, а также для распознавания образов. Без ее решения невозможен переход от концепции BigData к индустрии знаний. В этом прежде всего состоит преимущество стохастики по сравнению с современными западными технологиями.

Другим важнейшим достоинством стохастики является обеспечение гарантированной защиты данных и знаний в процессе их передачи, хранения и обработки, а также выполнения программ [2, 7, 11]. Это достигается за счет того, что вся информация циркулирует, а также обрабатывается в вычислительной среде компьютера в стохастически преобразованном, защищенном виде. И, наконец, стохастика, как будет показано ниже, совместима с нанотехнологией, что позволяет реализовать в ближайшем будущем интеллектуальные квантовые компьютеры и на этой основе – информационные проекты национального значения.

Литература

1. Черняк Л. Большие Данные — новая теория и практика // Открытые системы №10, 2011.
2. Насыпный В.В. Защищенные стохастические системы // Открытые системы №3, 2004.
3. Насыпный В.В. Распознавание и понимание смысла речи на основе стохастики в шумах. М.: Прометей, 2010. – 139 с.
4. Насыпный В.В. Развитие теории построения открытых систем на основе информационной технологии искусственного интеллекта. М.: Воениздат, 1994. - 248с.
5. Насыпный В.В., Насыпная Г.А. Способ синтеза самообучающейся системы извлечения знаний из текстовых документов для поисковых систем. Патент РФ №2273879, номер международной заявки PCT/RU02/00258, дата подачи 28 мая 2002.

6. Насыпный В.В., Насыпная Г.А. Способ синтеза самообучающейся аналитической вопросно-ответной системы с извлечением знаний из текстов, заявка на патент №2007120344/09 от 06.08.2007. Получено решение на выдачу патента на изобретение от 21.07.2008.
7. Насыпный В.В. Система с абсолютной стойкостью // Открытые системы №9, 2005.
8. Насыпный В.В., Насыпная Г.А. Система распознавания, понимания смысла, анимационного моделирования и синтеза речи на основе стохастической информационной технологии. М.: Прометей, 2008. – 76 с.
9. Искусственный интеллект. Справочник. Кн. 2. Модели и методы. Под ред. Поспелова Д.А. М.: Радио и связь, 1990. - 303 с.
10. Halsall F. Data communications computer networks and osi.Addison-wesley publishing company, 1988. - 973 с.
11. Насыпный В.В. Способ комплексной защиты распределенной обработки информации в компьютерных системах и система для осуществления способа. Патент РФ №2259639, номер международной заявки PCT/RU /00272, дата подачи 28.10.2003 г.
12. Насыпный В.В., Насыпная Г.А. Метод семантической связи текста с трехмерной графикой. – М.: Прометей, 2007. – 27 с.

2. Интеллект

От больших данных к индустрии знаний

В настоящее время происходит переход от широко распространенных информационных систем, включая Интернет, к интеллектуальным системам, основанным на знаниях. Как известно, в отличие от данных знания характеризуются определенным комплексом свойств и прежде всего активностью, новизной, эффективностью. Активность связана с возможностью автоматической генерации при актуализации знаний определенных информационных и аналитических процессов, направленных на обработку данных. Новизна характеризует содержание в знаниях новых сведений, не известных ранее конкретным пользователям системы в некоторой проблемной области. Эффективность знаний определяется возможностью реализации интеллектуальных процессов, позволяющих достичь конкретной цели или связанных с порождением новых знаний.

Основу интеллектуальных систем составляют базы знаний, в которых используется аппарат искусственного интеллекта, включающий семантические сети, фреймы, правила продукций, предикаты первого порядка и другие формализмы. Важнейшими элементами указанных систем являются также подсистемы логического вывода, интерпретации знаний, ввода-вывода полученных результатов. Системы, основанные на знаниях, широко используются для аналитической обработки информации, в

интеллектуальных поисковых системах, для распознавания и понимания смысла сенсорной информации.

На основе указанных систем строятся интеллектуальные сети, обеспечивающие эффективную обработку данных и знаний в интересах решения конкретных, достаточно сложных научно-технических и других задач, которые невозможно решить в рамках чисто информационных технологий.

В работах [4-6] предложены способы извлечения знаний из произвольной текстовой информации для реализации аналитических функций индукции, дедукции, сравнения, обобщения, аналогии, определения и других.

На основе указанных систем строятся базовые элементы интеллектуальной сети. Эта сеть составляет основу индустрии знаний, которая позволяет поднять на качественно новый уровень процессы управления обществом и производственной сферой, а также внести существенный вклад в ускоренное развитие науки и новых технологий.

Важнейшей проблемой развития индустрии знаний является их автоматическое приобретение путем обработки текстовой и сенсорной информации, в которой, прежде всего, выделяют естественную речь и видеоинформацию. Основой для решения этих задач является создание самообучающихся систем извлечения знаний из текстов, а также систем распознавания и понимания смысла речи и видеоинформации [3 - 6].

Решение отмеченных задач особенно актуально на современном этапе создания индустрии знаний, поскольку эти задачи напрямую связаны с реализацией в компьютерах функции понимания смысла.

Проблема понимания смысла при создании индустрии знаний

Важнейшей нерешенной проблемой BigData при создании автоматических (способных функционировать без участия человека) систем ввода и обработки текстовой и сенсорной информации является понимание смысла.

В современных системах обработки изображений создатели ограничиваются символьным представлением (описанием) отдельных элементов изображения, не затрагивая семантический (смысловой) уровень. При этом в создаваемых речевых технологиях проблема понимания смысла рассматривается как далекая перспектива, а вопрос о понимании смысла другой сенсорной информации (тактильной, связанной с обонянием, осязанием и др.) в настоящее время вообще не ставится разработчиками автоматизированных систем.

Вместе с тем, совершенно очевидно, что без решения проблемы понимания смысла дальнейшее развитие систем обработки текстовой и сенсорной информации не имеет прикладной перспективы.

Таким образом, для перехода от BigData к индустрии знаний необходимо решение проблемы понимания смысла текстовой и сенсорной информации.

Отметим, что под пониманием смысла поступающих знаний и сенсорной информации подразумевается способность их интерпретации (представления) с использованием иных терминов той же самой знаковой системы или какой-либо другой (прежде всего языковой).

Так, например, понимание смысла некоторого высказывания эквивалентно его переформулировке с использованием других терминов (иных слов) с полным сохранением смысла.

Создание полноценных систем понимания смысла текстов, речи и изображений невозможно без реализации функции автоматического самообучения при извлечении знаний из информационных сообщений и требует обеспечения возможности формирования нового знания и органичного (автоматического) дополнения этим знанием соответствующей опорной базы [2 - 6].

Результатом пополнения базы знаний системы является формирование концептуального описания проблемной среды («картины мира»), включающей объекты, субъекты, их классификацию, свойства, связи, соотношения, взаимодействия и др. На основе этой информации в интеллектуальной поисковой системе [5] обеспечивается автоматическое формирование смысла поступающих текстовых сообщений. Например, сообщение «футболист бежит по полю» автоматически переводится в сообщение «человек перемещается на местности». Для этого используется система классификации и «картина мира». На языке полученных классов объектов и отношений между ними система понимает смысл действий футболиста из первого сообщения и может его представить в виде второго сообщения на языке классов объектов и отношений между ними.

При этом реализация функции автоматического непрерывного формирования «картины мира» сопряжена с обязательным выполнением индуктивного логического вывода на больших пространствах знаний и данных. Именно логический вывод позволяет реализовать функции интеллектуальных систем, связанные с их самообучением путем формирования новых знаний, на основе смыслового содержания поступающей информации.

Отметим, что существующие информационные технологии не позволяют реализовать смысловую обработку ни текстов, ни изображений, ни речевых сообщений в силу не способности решить задачу индуктивного логического вывода на больших пространствах знаний и данных.

Однако до настоящего времени все попытки разработчиков интеллектуальных систем добиться осуществления индуктивной обработки информации наталкивались на проблему «комбинаторного взрыва», автоматически возникающую при попытках обработать в реальном времени соответствующие объемы информации методами перебора. Это не позволяет создать аппарат смысловой обработки текстовой, аудио- и видеоинформации.

Литература

1. Черняк Л. Большие Данные — новая теория и практика // Открытые системы №10, 2011.
2. Насыпный В.В. Защищенные стохастические системы // Открытые системы №3, 2004.
3. Насыпный В.В. Распознавание и понимание смысла речи на основе стохастики в шумах. М.: Прометей, 2010. – 139 с.
4. Насыпный В.В. Развитие теории построения открытых систем на основе информационной технологии искусственного интеллекта. М.: Воениздат, 1994. - 248с.
5. Насыпный В.В., Насыпная Г.А. Способ синтеза самообучающейся системы извлечения знаний из текстовых документов для поисковых систем. Патент РФ №2273879, номер международной заявки PCT/RU02/00258, дата подачи 28 мая 2002.

3. Логика

Стохастика как технология индуктивного логического вывода

Стохастика позволила адаптировать вычислительную среду современных компьютеров к обработке знаний и реализовать индуктивный вывод с использованием новых, непереборных методов обработки информации [4 - 6].

В основу предложенного метода положена единая процедура стохастического преобразования алфавитно-цифровой информации, описывающей семантическую сеть, которая включает фреймы и правила продукций. При этом обеспечивается непосредственное отображение семантической сети и правил продукций в вычислительную среду компьютера. Информационные единицы и связи будут взаимно однозначно поименованы стохастическими индексами и физически представлены в памяти компьютера стохастическими кодами. Достигается возможность эффективной реализации на базе сформированных индексов и кодов как процедур произвольного доступа к информации, так и логических операций на сети.

Это обеспечивает соответствие между произвольной символьной конструкцией и ее стохастическим индексом. Полученные уникальные индексы имеют двойственный характер: с одной стороны, они являются именем указанных символьных конструкций, с другой, - они определяют адрес, по которому необходимо произвести обращение к другим элементам знаний, логически (семантически) связанным с исходной символьной конструкцией [4].

При этом в процессе формирования индекса с помощью стохастической хэш-функции отображаются имеющиеся между символьными элементами связи типа «часть-целое», «род-вид», «причины-следствия», «условия-

заключения», «определения» и др. Так, например, при создании индекса словосочетаний используется индекс отдельных слов. Формирование стохастического индекса предиката производится на основе входящих в него индексов словосочетаний.

Индекс предложения реализуется с использованием стохастических индексов словосочетаний, предикатов, входящих в данное предложение и т.д.

На основе информации о составных элементах каждого индекса автоматически формируются новые знания о том, в какие индексы по критерию «часть-целое» и «род-вид» входит каждый элемент. Это позволяет в режиме активизации индексной информации путем реализации функций самообучения и автоматического формирования новых знаний описывать в индексной форме все возможные прямые логические связи исходного элемента с другими элементами на множестве пространства поиска [4].

В качестве этих элементов могут быть слова, словосочетания, предикаты, предложения, правила продукций и другие формы представления знаний.

Таким образом, после реализации описанного режима самообучения и автоматического получения индексных форм и логических связей над множеством элементов семантической сети или правил продукций формируется уровень метазнаний [4].

Указанный уровень метазнаний в виде некоего виртуального информационного поля определяет все возможные траектории логического вывода на каждом его шаге, отбирая только семантически связанные символьные конструкции (слова, словосочетания, предикаты, правила продукций и др.) и элементы знаний.

За счет этого устраняется необходимость полного перебора на каждом шаге логического вывода и снимается проблема «комбинаторного взрыва». Каждая траектория логического вывода содержит в качестве своих элементов только неповторяющиеся символьные конструкции знаний. Повторение символьных конструкций приводит к образованию циклов, что свидетельствует о необходимости корректировки баз знаний с целью устранения указанных повторов [4].

При реализации логического вывода на множестве N элементов знаний любой его траектории требуется обработать не более M << N символьных элементов знаний, представленных уникальными стохастическими индексами. Следовательно, время логического вывода при использовании описанного метода, основанного на стохастической информационной технологии, будет линейно зависеть от числа M логически или семантически связанных символьных конструкций (слов, словосочетаний, предикатов, элементов семантической сети или правил продукций) [4].

Количество M элементов, применяемых в процессе построения любой траектории логического вывода будет значительно ниже, чем максимальное число N этих элементов в пространстве поиска требуемых символьных конструкций.

Отметим, что предложенные методы логического вывода на основе стохастической информационной технологии позволяют выбрать минимально допустимую и наиболее вероятную траекторию логического вывода на любом множестве семантически связанных символьных конструкций и построить метаправила для обеспечения обработки знаний в заданное время. Это позволяет создать на базе существующих компьютеров эффективные интеллектуальные системы, работающие в любом поисковом пространстве без сужения множества возможных гипотез лингвистического анализа и смыслового поиска в реальном масштабе времени. Указанные системы описаны в патентах [5, 6].

Кроме этого предложенный метод логического вывода позволяет реализовать новые технологии распознавания речевых сообщений и изображений, основанные на семантическом анализе и логической обработке знаний. Эти методы являются универсальными. Они дают возможность синтезировать более эффективные и достоверные технологии распознавания речевых сообщений от неизвестного диктора на неограниченном объеме словаря. При этом пиксельное представление речевого сигнала преобразуется в семантические образы, которые с помощью знаний описываются как понятия, связанные с артикуляционной и акустической классификацией сигнала. Это позволяет с высокой достоверностью, приближающейся к 100%, распознавать звуки, фонемы, слоги и слова, то есть реализовывать эффективный фонетический анализ речи и перевести звуковые образы в достоверные лексические элементы [3].

При распознавании видеоинформации методы, основанные на знаниях и семантическом анализе изображений, позволяют эффективно обрабатывать информацию на всех трех уровнях представления изображений: пиксельном, пиксельно-контурном и уровне символьного описания полученного изображения.

После этого как для речевых сообщений, так и для изображений включаются уровни лингвистического и семантического анализа с использованием методов индуктивного логического вывода и обработки смысловых конструкций. В результате осуществляется семантическая классификация понятий и формирования концептуального описания («картины мира»), которые представлены выше. Впервые достигается возможность достоверного распознавания слитных речевых сообщений от неизвестного диктора и различных видеоизображений (двухмерных, трехмерных), включая видеосъемку, в реальном времени.

При этом ввод и обработка семантики поступающих в систему текстов речевых сообщений или изображений активизирует специальную процедуру логического вывода, позволяющую извлекать знания (в том числе новые), проверять их корректность и органично вписывать в состав опорной базы знаний.

Литература

1. Черняк Л. Большие Данные — новая теория и практика // Открытые системы №10, 2011.
2. Насыпный В.В. Защищенные стохастические системы // Открытые системы №3, 2004.
3. Насыпный В.В. Распознавание и понимание смысла речи на основе стохастики в шумах. М.: Прометей, 2010. – 139 с.
4. Насыпный В.В. Развитие теории построения открытых систем на основе информационной технологии искусственного интеллекта. М.: Воениздат, 1994. - 248с.
5. Насыпный В.В., Насыпная Г.А. Способ синтеза самообучающейся системы извлечения знаний из текстовых документов для поисковых систем. Патент РФ №2273879, номер международной заявки PCT/RU02/00258, дата подачи 28 мая 2002.
6. Насыпный В.В., Насыпная Г.А. Способ синтеза самообучающейся аналитической вопросно-ответной системы с извлечением знаний из текстов, заявка на патент №2007120344/09 от 06.08.2007. Получено решение на выдачу патента на изобретение от 21.07.2008.
7. Насыпный В.В. Система с абсолютной стойкостью // Открытые системы №9, 2005.
8. Насыпный В.В., Насыпная Г.А. Система распознавания, понимания смысла, анимационного моделирования и синтеза речи на основе стохастической информационной технологии. М.: Прометей, 2008. – 76 с.
9. Искусственный интеллект. Справочник. Кн. 2. Модели и методы. Под ред. Поспелова Д.А. М.: Радио и связь, 1990. - 303 с.
10. Halsall F. Data communications computer networks and osi.Addison-wesley publishing company, 1988. - 973 с.
11. Насыпный В.В. Способ комплексной защиты распределенной обработки информации в компьютерных системах и система для осуществления способа. Патент РФ №2259639, номер международной заявки PCT/RU /00272, дата подачи 28.10.2003г.
12. Насыпный В.В., Насыпная Г.А. Метод семантической связи текста с трехмерной графикой. – М.: Прометей, 2007. – 27с.

4. Квантовый компьютер с интеллектом

Квантовая технология и стохастика

Одним из достижений быстроразвивающейся нанотехнологии является ясно наметившаяся перспектива создания квантового компьютера [13]. Как известно, разрабатываемые в нанотехнологии квантовые компьютеры в

отличие от существующих компьютеров могут перерабатывать информацию, исходя из представления о так называемом квантовом бите (кубите или нанобите).

Рис. 0.1. Схематическая структура квантового компьютера.

При этом L кубитов соответствует числу $N = (2^L-1)$, что позволяет вычислителям работать с очень большими числами или большими объемами данных. На рис. 1 представлена схематическая структура квантового компьютера. Основной частью его является квантовый регистр – совокупность некоторого числа L кубитов. Далее следуют блок ввода информации, блок квантовых вычислений, блок измерения состояния кубитов. Для управления этими блоками применяется классический управляющий компьютер, включающий генератор воздействующих на кубиты импульсов [14].

При реализации в нанотехнологиях база данных может состоять из N сообщений, представленных $N = 1 \div (2^L-1)$ состояниями квантового регистра из L кубитов. При этом любое из конкретных сообщений с использованием алгоритма Гровера [14] может быть найдено по значению его кода длиной в Lкубитов. При вводе информации в квантовый компьютер состояние входного регистра преобразуется в соответствующую комбинацию. В таком виде информация далее подвергается воздействию квантового процессора, выполняющего последовательность квантовых логических операций. В результате преобразований исходное квантовое состояние соответствующим образом изменяется, что фиксируется с помощью измерения состояния кубитов на выходе компьютера [14].

Совокупность всех возможных операций на входе данного компьютера, формирующих исходное состояние, а также воздействий, осуществляющих унитарные локальные преобразования, соответствующие алгоритму вычисления, играют здесь ту же роль, что и программное обеспечение в классическом компьютере, реализующем управляющие функции [14].

Описываемый квантовый компьютер предназначен, в основном, для вычислительных функций с большими числами, которые трудно реализуются в классическом компьютере. Вопрос о построении с помощью квантового

компьютера интеллектуальных систем в соответствии с описанными выше требованиями индустрии знаний пока не стоит. Однако применение стохастики, как будет показано ниже, позволяет решить эту проблему в самое ближайшее время.

В стохастике любой элемент данных или сообщение представлены стохастическими индексами или кодами длиной L-бит. Способ автоматического формирования уникальных стохастических индексов из каждой символьной конструкции U приведен в работе [4]. При этом значение каждого индекса равновероятно распределено в интервале $1 \div (2^L-1)$.

Общее число элементов данных или знаний N также, как и в нанотехнологии, определяется из следующей формулы: $N = 1 \div (2^L-1)$. Поэтому основная идея использования стохастики в квантовых компьютерах заключается в замене кубитов на стохастические индексы символьных конструкций, которые можно назвать «стохастическими кубитами». Поскольку стохастика использует для логической обработки простейшие процедуры типа произвольного доступа к данным, булевой алгебры и логической обработки индексов, то такой подход позволит реализовать на основе стохастических кубитов все описанные выше функции интеллектуальной обработки данных и знаний, представленных в работах [2-6].

В качестве примера рассмотрим порядок реализации в квантовом компьютере на основе стохастики базы данных и знаний, приведенной в работе [4].

База данных в стохастике содержит концептуальную часть, описывающую в виде семантической сети «картину мира» в заданной предметной области. Фактуальная часть базы данных, объем которой может на несколько порядков превышать объем семантической сети, описывает свойства конкретных объектов данного класса или типа в виде отношений с другими объектами предметной области. При этом концептуальная часть размещается в обычном компьютере, а фактуальная часть может загружаться в квантовый компьютер. С помощью логического вывода на семантической сети и булевой алгебры формируются стохастические индексы доступа {}, идентичные кубитам в квантовом компьютере. После этого, используя алгоритм Гровера [14], производится доступ к нужному сообщению фактуальной части базы данных. Затем с помощью логического вывода и базы знаний выбираются все кубиты, относящиеся к выбранному сообщению. Это обеспечивает эффективную обработку в реальном времени больших данных. Например, характеризуемые значением $N= 2^{100}-1 = 10^{30}$ сообщений и более. За счет этого реализуются комбинированные интеллектуальные системы, включающие нанотехнологию и стохастику, обеспечивающие обработку больших данных в реальном времени.

Рассмотрим теперь порядок реализации в квантовом компьютере логического вывода в больших базах данных и знаний на основе стохастики. Как было отмечено выше, эта процедура является базовой для создания

интеллектуальных систем, обеспечивающих переход от больших данных к индустрии знаний.

Метод и технология логического вывода в квантовом компьютере на основе стохастики

Известно, что одной из основных проблем, возникающих при разработке интеллектуальных систем, является определение возможных траекторий логического вывода на множестве правил продукций, данных и знаний. Это обеспечит отход от переборного метода логического вывода, который в больших базах знаний приводит к «комбинаторному взрыву».

Для решения этой проблемы в [4] предложен метод определения возможных траекторий, поиска целей и предварительного выбора кратчайшего пути логического вывода, основанный на построении сети правил продукций и оперативного режима их обработки с использованием стохастических индексов и кодов. За счет применения стохастической информационной технологии, как будет показано ниже, проблема комбинаторного взрыва полностью снимается, поскольку время логического вывода будет линейно зависеть от числа правил продукций, непосредственно задействованных в обработке. Отметим, что наполнения базы знаний правилами продукций может производиться в автоматическом режиме с использованием самообучающихся интеллектуальных систем извлечения знаний из текста, описанных в работах [3, 4]. Из изложенного следует, что вместо стохастических индексов в этих процедурах могут применяться стохастические кубиты, обрабатываемые в квантовом компьютере.

Для реализации указанной возможности каждое правило продукций подвергается стохастическому преобразованию и имеет следующий вид:

(1)

где - стохастические индексы предикатов условия,

- индекс предиката заключения правила.

Если предикат терминальный, то он связан посредством индекса с атомарной формулой вида , где - соответственно коды атрибутов u_i и их значений a_i фреймов или реляционных отношений, θ - арифметический оператор сравнения (≥, ≤, = и др.).

В процессе обработки терминального предиката , в квантовом компьютере по индексу производится доступ к соответствующему фрейму или реляционному отношению. При этом используется алгоритм Гровера [14]. Истинность или ложность предиката устанавливается путем сравнения кодов атомарной формулы, определяемой индексом с кодом фрейма или реляционного отношения. Сравнение индексов и кодов может

осуществляться в управляющем компьютере (рис.1) с использованием стохастического процессора, обеспечивающего выполнение логических и арифметических функций со стохастическими индексами и кодами без их расшифровки [2]. Реализация процедур обработки терминальных предикатов путем произвольного доступа по стохастическим индексам к соответствующим фреймам и реляционным отношениям базы данных и знаний квантового компьютера, а также выполнение функций сравнения стохастических индексов и кодов в стохастическом процессоре обеспечивает высокую эффективность обработки терминальных предикатов.

Эти параметры и значения после стохастической индексации записываются в соответствующие фреймы базы знаний.

Проблема устранения комбинаторного взрыва решается, как было отмечено выше, на основе автоматического формирования сети правил продукций системы логического вывода в режиме самообучения с использованием агрегативных свойств стохастических индексов. Эта система может быть реализована непосредственно в квантовом компьютере или в управляющем классическом компьютере (рис.1). В данном случае стохастический индекс каждого правила продукций вычисляется путем сложения по mod 2 индексов всех m предикатов каждого правила:

(2)

Таким образом, стохастический индекс правила связан со стохастическими индексами составляющих его предикатов как «целое-часть». На этой основе для построения сети правил продукций системы логического вывода, автоматически формируются новые знания - специальные сетевые фреймы, в которые включаются стохастические индексы правил, имеющих идентичные предикаты в условии или заключении.

Указанные сетевые фреймы формируются для каждого правила базы знаний. При этом индексы правил , имеющих предикат в условии или заключении, идентичный предикату , будут включены в один и тот же сетевой фрейм. Это обусловлено тем, что доступ к сетевым фреймам осуществляется по индексам предикатов . Поэтому все индексы правил , относящиеся к одному и тому же предикату, будут содержаться в одном и том же сетевом фрейме. Объединенные логическими связками ИЛИ они определяют альтернативные направления логического вывода.

Образованная сетевая структура позволяет реализовать прямую и обратную последовательности логического вывода, используя только связанные друг с другом идентичными предикатами правила продукций. Тем самым исключается необходимость перебора на всем множестве правил продукций при выполнении каждого очередного цикла логического вывода. В результате исключается главная причина комбинаторного взрыва. При этом за счет использования сетевых фреймов при построении траектории

логического вывода каждое правило обрабатывается не более одного раза. Поэтому при реализации любой возможной траектории логического вывода на множестве правил продукций может быть задействовано не более общего числа M<<N правил продукций в базе знаний. Поэтому время логического вывода с использованием предложенного метода на основе стохастической информационной технологии будет линейно зависеть от числа правил продукций, непосредственно входящих в траекторию логического вывода в базе знаний:

(3)

где - время логического вывода на множестве правил продукций,

а - время обработки правила продукций.

Отметим, что применение сетевых правил продукций позволяет осуществить выбор кратчайшего пути логического вывода.

Для предварительного выбора кратчайшего пути логического вывода в работе [4] предложен оперативный режим генерации дерева траекторий логического вывода. Он основан на применении процедур доступа по стохастическим индексам к сетевым фреймам и фреймам правил с отложенной обработкой терминальных предикатов (высказываний), требующих обращения к памяти квантового компьютера.

При реализации этого режима после выбора целевого правила на каждом последующем уровне j дерева траекторий из сетевых фреймов образуется логическое выражение, включающее индексы правил, заключения которых связаны идентичными предикатами с условиями правил (j-1)-го уровня. Каждому правилу (j-1)-го уровня в общем случае может соответствовать набор сетевых фреймов, содержащих индексы правил, заключения которых связаны с условием правила . При формировании логического выражения уровня индексы правил внутри сетевых фреймов, объединенные логическими связками ИЛИ, заключаются в скобки; сетевые фреймы, относящиеся к одному правилу (j)-го уровня, закрываются дополнительными скобками. В результате будет сформировано логическое выражение вида:

, (4)

которое поступает в дальнейшую обработку.

Как показано в работе [4], применение режима оперативной генерации траекторий логического вывода с помощью сетевых фреймов дает возможность предварительного определения кратчайшего пути логического вывода в виде одной из полученных составных продукций, обеспечивающей достижение цели логического вывода посредством обработки минимального числа терминальных предикатов.

Рассмотрим порядок реализации логического вывода, и построения необходимых цепочек правил продукций с использованием базы данных и знаний и алгоритма Гровера квантового компьютера. В результате описанного выше алгоритма логического вывода с использованием сетевых правил образуется некоторое дерево логического вывода. Допустим, что указанное дерево содержит три уровня логически связанных через сетевые фреймы правил продукций - . После генерации логического выражения на каждом уровне можем получить следующий результат:

(5)

Сформированное логическое выражение поступает в дальнейшую обработку для определения истинности предиката целевого правила продукций .

Покажем, что применение режима оперативной генерации траекторий логического вывода с помощью сетевых фреймов позволяет предварительно сформировать все необходимые цепочки логического вывода и определить его кратчайший путь. Под кратчайшим путем логического вывода будем понимать одну из возможных траекторий (цепочек правил), обеспечивающую достижение цели логического вывода посредством обработки минимального числа терминальных предикатов.

Для этого используем понятие составной продукции[9], которая представляет собой совокупность всех терминальных предикатов цепочки правил продукций дерева траекторий логического вывода, объединенных логическими связками $\wedge$ и обеспечивающих независимое определение истинности или ложности целевого предиката . Составная продукция имеет следующий вид:

, (6)

где - терминальные предикаты цепочки правил продукций;

- предикат заключения целевого правила продукций.

Легко видеть, что совокупность составных продукций может быть получена из выражения (5), сформированного в результате оперативной генерации дерева траекторий логического вывода после раскрытия скобок. По сути дела, составная продукция представляет собой новое знание, полученное из исходной цепочки правил путем формирования новой логической конструкции.

Для рассматриваемого примера из сформированного выражения можно получить четыре независимые цепочки логического вывода:

(7)

Из каждой цепочки правил продукций, исключив логически связанные через сетевые фреймы предикаты, получим следующие четыре составные продукции (P_{s1} , P_{s2} , P_{s3} и P_{s4}):

(8)

Определив число терминальных предикатов в каждой из составных продукций Psi, мы можем выделить составную продукцию с минимальным числом терминальных предикатов в условии. Очевидно, что эта составная продукция Ps2 и является прогнозируемым кратчайшим путем логического вывода, минимизирующим время определения истинности целевого предиката при обработке терминальных предикатов условия.

Таким образом, реализация логического вывода "вширь" с использованием режима оперативной генерации выражения (8) перед выполнением операций сопоставления с образцом, требующих для обработки терминальных предикатов доступа к базе данных и знаний квантового компьютера, выдает кратчайший путь логического вывода. При этом остальные составные продукций могут быть расположены в порядке возрастания числа терминальных предикатов для оптимизации последующей обработки дерева логического вывода, если прогнозируемый кратчайший путь не позволяет определить истинность целевого предиката .

Процедура синтеза составных продукций и предварительный выбор кратчайшего пути можно применить и при генерации прямой волны логического вывода. Очевидно, что сформировав и преобразовав логическое выражение, полученное в результате оперативной генерации прямой волны логического вывода, мы получим набор составных продукций {Psj}, каждая из которых содержит исходное правило и заканчивается одним из возможных целевых правил продукций. Это позволяет выбрать составную продукцию, содержащую минимальное число терминальных предикатов.

Реализация в квантовом компьютере оперативной генерации траекторий логического вывода с использованием сетевых фреймов дает возможность произвести предварительный выбор кратчайшего пути прямой и обратной последовательностей вывода, минимизирующих время последующей обработки терминальных предикатов при выполнении процедуры сопоставления с образцом. После этого производится реализация

описанных выше процедур обработки стохастических кубитов квантового компьютера с использованием алгоритма Гровера.

* * *

Таким образом, применение стохастики позволит в кратчайшие сроки пройти путь от теории и практики больших данных к индустрии знаний и нанотехнологии. Это даст возможность разработать принципиально новые технические системы практически во всех промышленных сферах, включая новые поколения компьютеров, суперинтеллектуальные защищенные системы в робототехнике, в авиакосмической, атомной промышленности и других высокотехнологичных областях. Кроме того, суперинтеллектуальные системы смогут играть все возрастающую роль при решении социальных проблем национального значения, а также в управлении обществом. Например, как показали исследования, на основе описанного подхода в ближайшие три-пять лет могут быть реализованы IT проекты, имеющие национальное значение, а именно:

1. Создание на Федеральном уровне интеллектуальной системы контроля аудио-, видео- и текстовой информации с распознаванием смысла и содержания действий, производимых контролируемыми объектами и субъектами. При этом обеспечивается автоматическое своевременное выявление и предупреждение противоправных действий в общественных местах, нарушений правил дорожного движения, положений техники безопасности на производстве и др.

2. Создание интеллектуальной системы предотвращения возможности использования государственных денежных средств в коррупционных схемах. Это обеспечивается на основе применения «зашифрованных электронных денег» и контроля финансовых потоков в государственных и коммерческих системах с гарантированной идентификацией и аутентификацией отправителей и получателей денежных средств, с определением в реальном времени цели поступления финансов и их использования в соответствии с требованиями закона.

3. Создание интеллектуальной системы защиты от распространения в торговой сети контрафактной продукции, прежде всего медикаментов, пищевой продукции, аудио- и видеоматериалов и других товаров на основе применения «зашифрованных электронных денег» и контроля траектории их прохождения для определения законности заключаемых договоров и сделок между разработчиками и торгующими организациями.

Литература

1. Черняк Л. Большие Данные — новая теория и практика // Открытые системы №10, 2011.

2. Насыпный В.В. Защищенные стохастические системы// Открытые системы №3, 2004.
3. Насыпный В.В. Распознавание и понимание смысла речи на основе стохастики в шумах. М.: Прометей, 2010. – 139 с.
4. Насыпный В.В. Развитие теории построения открытых систем на основе информационной технологии искусственного интеллекта. М.: Воениздат, 1994. - 248с.
5. Насыпный В.В., Насыпная Г.А. Способ синтеза самообучающейся системы извлечения знаний из текстовых документов для поисковых систем. Патент РФ №2273879, номер международной заявки PCT/RU02/00258, дата подачи 28 мая 2002.
6. Насыпный В.В., Насыпная Г.А. Способ синтеза самообучающейся аналитической вопросно-ответной системы с извлечением знаний из текстов, заявка на патент №2007120344/09 от 06.08.2007. Получено решение на выдачу патента на изобретение от 21.07.2008.
7. Насыпный В.В. Система с абсолютной стойкостью // Открытые системы №9, 2005.
8. Насыпный В.В., Насыпная Г.А. Система распознавания, понимания смысла, анимационного моделирования и синтеза речи на основе стохастической информационной технологии. М.: Прометей, 2008. – 76 с.
9. Искусственный интеллект. Справочник. Кн. 2. Модели и методы. Под ред. Поспелова Д.А. М.: Радио и связь, 1990. - 303 с.
10. Halsall F. Data communications computer networks and osi. Addison-wesley publishing company, 1988. - 973 с.
11. Насыпный В.В. Способ комплексной защиты распределенной обработки информации в компьютерных системах и система для осуществления способа. Патент РФ №2259639, номер международной заявки PCT/RU /00272, дата подачи 28.10.2003г.
12. Насыпный В.В., Насыпная Г.А. Метод семантической связи текста с трехмерной графикой. – М.: Прометей, 2007. – 27с.
13. Кобаяси Н. Введение в нанотехнологию / Н.Кобаяси. – Пер. с японск. – М.: БИНОМ. Лаборатория знаний, 2007. – 134 с.
14. Валиев К.А., Кокин А.А. Квантовые компьютеры: надежды и реальность. – Москва-Ижевск: НИЦ «Регулярная и хаотическая динамика», 2004, 329 стр.

5. Автоматическое понимание смысла и реферирование текста на основе стохастики

Обработка текстов на основе стохастики

Основным при разработке методов понимания и реферирования неструктурированной текстовой информации является использование лингвистического и семантического анализа. Эти виды анализа требуют доступа и обработки к большим объемам знаний (например, к «картине

мира») и решения проблемы BigData. Применение стохастической технологии в отличие от существующих систем позволяет решить эту проблему без возникновения эффекта комбинаторного взрыва [2].

Это обусловлено тем, что современные методы обработки символьной информации, представляющей знания, используют вычислительные алгоритмы над строками символов, которые реализуются по известным алгоритмам машины Тьюринга. Указанные алгоритмы, предназначенные для реализации вычислительных функций, не ориентированы на эффективное выполнение операций логического вывода в пространстве семантической сети и при использовании правил продукций [2].

Данное положение объясняется, прежде всего, тем, что в существующих алгоритмах поиска имя (наименование) символьной конструкции и ее физический адрес в памяти компьютера суть два различных элемента. Поэтому на каждом шаге логического вывода сначала производится поиск нужной символьной конструкции путем перебора на множестве всех возможных ее значений, а затем определяется адрес с целью дальнейшей обработки. По данному адресу выбирается необходимая символьная информация для реализации следующего шага логического вывода. При этом поиск очередной логически связанной символьной конструкции также производится путем перебора.

Известно [1], что для поиска одной произвольной записи среди неупорядоченного множества элементов теоретический предел объема вычислений составляет $N\log_2 N$, где N – число записей в массиве, по которому выполняется поиск. Поэтому при выполнении поисковых операций для каждого элемента поискового запроса (например, каждого слова поискового запроса) на большом массиве записей получается огромный объем вычислений. Это не позволяет в реальном масштабе времени проводить поиск на больших объемах информации (проблема BigData).

В результате время логического вывода увеличивается по экспоненциальному закону в зависимости от N, где N – число возможных символьных конструкций знаний [2].

Покажем важность решения проблемы комбинаторного взрыва на примере попыток создания ведущими корпорациями мира (Microsoft, Google и др.) интеллектуальных поисковых систем с семантическим анализом текста. Отметим, что данные корпорации в настоящее время разработали эффективные системы поиска по ключевым словам.

Для сокращения количества вычислений на этапе поиска во всех поисковых системах используется предварительная обработка текстовой информации – так называемая индексация. В процессе такой обработки для каждого слова индексируемых текстов указывается его уникальное значение (индекс), а также наборы указателей на тексты и позиции в текстах, в которых это слово встречалось. Такой формат представления информации называется «обратный индекс». Это обуславливает необходимость поиска каждого слова во множестве $N = 10^5$ символьных конструкций, что определяется средним объемом словаря, в котором производится поиск ключевых слов. При этом

высокая производительность современных компьютеров и методов распараллеливания поисковых операций позволяет избежать существенного замедления процесса поиска и реализует его в масштабе реального времени.

За счет использования обратного индекса возможно быстро находить тексты, в которых имеется группировка слов поискового запроса. Поэтому при реализации традиционного поиска по ключевым словам эффекта «комбинаторного взрыва» не возникает по причине использования обратного индекса и предварительной индексации текстов.

Для перехода к процедуре семантического анализа текстов с целью понимания смысла возникает необходимость логического вывода на множестве всех возможных понятий словаря ($N = 10^5$), словосочетаний, максимальное число которых $N = 10^{10}$, а также предикатов, описывающих «картину мира» (эволюционную предметную область) проиндексированных текстов. При этом в процессе построения траектории логического вывода на множестве указанных символьных конструкций, как правило, требуется выполнить несколько шагов логического вывода.

Это связано с необходимостью неоднократного перебора на множестве N символьных элементов, который неминуемо приводит к эффекту комбинаторного взрыва.

Например, при классификации понятий, словосочетаний и предикатов «картины мира» для их семантического анализа требуется произвести логический вывод на множестве 10^5 понятий и дефиниций толкового словаря. При этом необходимо выполнить два этапа логического вывода.

На первом этапе для классификации понятий необходимо осуществить не менее $N^{(1)} = 10^5 \cdot 10^5 = 10^{10}$ обращений к понятиям и дефинициям толкового словаря. На втором этапе при классификации словосочетаний и предикатов требуется не менее $N^{(2)} = 10^{10} \cdot 10^5 = 10^{15}$ обращений к толковому словарю. Общее число обращений к словарю для классификации понятий и словосочетаний будет равно $N = N^{(1)} + N^{(2)} > 10^{15}$.

Отметим, что минимально возможное время логической обработки такого количества $N = 10^{15}$ понятий и их дефиниций для современных компьютеров с быстродействием $V = 10^9$ (оп/с) примерно равно $T = 10^6$ (с). Это соответствует приблизительно 12 суткам. При построении интеллектуальных систем понимания текста и распознавания речи на базе Интернет классификация словосочетаний и предикатов предметной области должна осуществляться регулярно в реальном времени, соответствующем частоте обновления информации в проиндексированных текстах на сайтах системы. Поэтому для традиционных технологий обработки символьной информации указанная классификация, не реализуемая в реальном масштабе времени, не обеспечивает корректность семантического анализа.

Даже однократный перебор на множестве всех возможных словосочетаний, максимальное число которых $N = 10^{10}$, требует не менее $T = 10$(с) времени, что также не соответствует требованиям к быстродействию интеллектуальной поисковой системы.

Таким образом, при логическом выводе на реальных множествах символьных конструкций неструктурированной текстовой информации число переборов увеличивается по экспоненциальному закону. В этом случае возникающий комбинаторный взрыв делает невозможным создание интеллектуальных систем распознавания речи с семантическим анализом текста и пониманием смысла в реальном масштабе времени на основе использования традиционной технологии поиска [2].

При реализации лингвистического анализа текстов, на основе которого реализуется понимание смысла, актуальность задачи «комбинаторного взрыва» также существенно возрастает. Это обусловлено комплексом причин.

1. Многозначные слова имеют различные морфологические индексы. Это обусловлено тем, что разные части речи имеют различные правила словообразования, которые должны учитываться при поиске. Особенно это актуально для агглютинативных языков (английский и др.).

2. Синтаксический анализ предложения базируется на построении и проверки гипотез синтаксического разбора слов в предложении анализируемого текста. Для синтаксического анализа предложений используются множества правил лингвистического анализа. Данные правила объединяются в сложную иерархическую систему групп знаний и логического вывода. Смысловая неоднозначность омонимов приводит к возможности появления нескольких корректных вариантов синтаксического разбора одного и того же предложения. Синтаксический анализ приводит, как правило, к нескольким гипотезам разбора предложения. Для устранения комбинаторного взрыва при использовании традиционных технологий используются вручную формируемые правила группировки лингвистических конструкций и правил лингвистического анализа. Это обеспечивает возможность создания предметно-ориентированных семантических систем, но он не способствует созданию универсальных систем логического вывода, которые обеспечивают необходимую полноту и корректность поиска в любой проблемной области. Реализация корректного синтаксического анализа невозможна без одновременного применения методов семантического анализа, связанного с использованием картины мира, что требует решения проблемы BigData [2, 12].

Указанная проблема решается с помощью стохастической информационной технологии. Сущность новой технологии заключается в стохастическом (случайном) преобразовании символьных конструкций (формульных выражений), правил продукций, элементов семантической сети, слов, словосочетаний, предикатов, названий, предложений, абзацев и других текстовых конструкций в уникальные стохастические индексы (коды заданной длины) [2].

Это обеспечивает взаимооднозначное соответствие между произвольной символьной конструкцией и ее стохастическим индексом. При этом полученные уникальные индексы имеют двойственный характер: с одной стороны, они являются именем указанных символьных конструкций, с другой, - они определяют адрес, по которому необходимо произвести

обращение к другим элементам знаний, которые непосредственно логически (семантически) связаны с исходной символьной конструкцией [2].

При этом в процессе формирования индекса с помощью стохастической хэш-функции отображаются имеющиеся между символьными элементами связи типа «часть-целое» или «род-вид». Так, например, при создании индекса словосочетаний используется индекс отдельных слов. Формирование стохастического индекса предиката производится на основе входящих в него индексов словосочетаний и отдельных слов.

Индекс предложения реализуется с использованием стохастических индексов словосочетаний, предикатов, входящих в данное предложение и т.д. При этом, за счет свойств стохастического преобразования, обеспечивается уникальность каждого полученного индекса со сколь угодно малой, заданной вероятностью коллизий [2].

На основе информации о составе каждого индекса (из каких индексов он образован) автоматически формируются новые знания о том, в какие индексы по критерию «часть-целое» и «род-вид» входит каждый элемент. Это позволяет в режиме активизации индексной информации путем реализации функций самообучения и автоматического формирования новых знаний описывать в индексной форме все возможные прямые логические связи исходного элемента с другими элементами на множестве пространства поиска.

В качестве этих элементов могут быть слова, словосочетания, предикаты, предложения, правила продукций и другие формы представления знаний.

Таким образом, после реализации описанного режима самообучения и автоматического получения индексных форм и логических связей над множеством элементов семантической сети или правил продукций формируется уровень метазнаний.

Указанный уровень метазнаний в виде некоего виртуального информационного поля определяет все возможные траектории логического вывода на каждом его шаге, отбирая только семантически связанные символьные конструкции (слова, словосочетания, предикаты, правила продукций и др.) и элементы знаний.

За счет этого устраняется необходимость полного перебора на каждом шаге логического вывода и снимается проблема комбинаторного взрыва. При этом каждая траектория логического вывода содержит в качестве своих элементов только неповторяющиеся символьные конструкции знаний. Повторение символьных конструкций приводит к образованию циклов, что свидетельствует о необходимости корректировки баз знаний с целью устранения указанных повторов [2].

Поэтому при реализации любой траектории логического вывода требуется обработать не более $M<<N$ символьных элементов знаний, представленных уникальными стохастическими индексами. Следовательно, время логического вывода при использовании описанного метода,

основанного на стохастической информационной технологии, будет линейно зависеть от числа M логически или семантически связанных символьных конструкций (слов, словосочетаний, предикатов, элементов семантической сети или правил продукций) [2].

Количество M элементов, применяемых в процессе построения любой траектории логического вывода, будет значительно ниже, чем максимальное число N этих элементов в пространстве поиска требуемых символьных конструкций. Например, при описанной выше процедуре классификации понятий и словосочетаний предметной области с использованием толкового словаря, максимальная траектория логического вывода содержит не более $M = 10^3$ семантически связанных по критерию «род-вид» символьных элементов словаря. Логический вывод производится в текстовых структурах словаря Ожегова при определении множества всех понятий, относящихся к классу «место». Поэтому при классификации понятий и словосочетаний предметной области текста с использованием стохастической информационной технологии требуется не более $N = 10^5 \cdot 10^3 = 10^8$ обращений к толковому словарю. При этом минимально возможное время логического вывода $T = 10^{-1}$ (с).

Отметим, что предложенный метод логического вывода на основе стохастической информационной технологии позволяет выбрать минимально допустимую и наиболее вероятную траекторию логического вывода на любом множестве семантически связанных символьных конструкций и построить метаправила для обеспечения обработки знаний в заданное время. Это дает возможность создать на базе существующих компьютеров эффективные интеллектуальные системы, работающие в любом поисковом пространстве без сужения множества возможных гипотез лингвистического анализа и смыслового поиска в реальном масштабе времени. Указанные системы описаны в патентах [3, 4]. Учитывая важность обеспечения эффективного логического вывода на множестве правил продукций при выполнении комплексной обработки текстовой информации с использованием баз знаний рассмотрим более подробно метод лингвистического анализа на основе стохастической информационной технологии.

Автоматическое определение морфологических, синтаксических и семантических характеристик слов

Для реализации процесса полного лингвистического анализа и понимания неструктурированного текста предложен программный комплекс интеллектуальных систем. Этот комплекс включает самообучающуюся аналитическую систему с извлечением знаний из текстов, а также интеллектуальные системы анализа и понимания смысла текстовой информации [3, 4]. В данном разделе рассмотрим концептуальные основы построения самообучающейся аналитической системы, которая предназначена, прежде всего, для лингвистического и прагматического

(семиотического) анализа с целью определения смысла неструктурированного текста.

Отметим, что понимание смысла речи проводится на двух уровнях – семантическом и прагматическом [1]. При этом понимание смысла отдельных членов, словосочетаний и предикатов предложения производится на семантическом уровне, понимание смысла предложений, абзацев и других текстовых структур выполняется на прагматическом уровне.

Для этого требуется эффективная интеллектуальная обработка с использованием больших объемов знаний и реализации логического вывода в реальном масштабе времени в режиме BigData. Отметим, что современные интеллектуальные системы не обеспечивают решения указанных задач ввиду эффекта комбинаторного взрыва. Как показано в работе [2], эта задача успешно решается на основе стохастической информационной технологии.

В данном разделе описан порядок построения и применения самообучающихся интеллектуальных аналитических систем с извлечением знаний из текстов для понимания смысла текста. Эти изделия подробно описаны в [3, 4].

Как было отмечено выше, указанные системы создаются на основе стохастической информационной технологии. Цель - построение на базе современного компьютера (машины Тьюринга) нового виртуального компьютера для эффективной лингвистической, семантической и логической обработки текстов.

Выбор тематики аналитических систем определяется содержанием неструктурированной текстовой информации, предоставленной для смыслового анализа. При этом аналитические функции, реализуемые в системе, которые связаны с индуктивным и дедуктивным логическим выводом, аналогией, обобщением, сравнением и др., широко применяются в ходе семантического и прагматического анализа полученного текста. Отметим, что при самообучении системы происходит формирование «картины мира» и системы семантической классификации понятий, словосочетаний и предикатов, входящих в состав картины мира, без которых невозможен полноценный семантический анализ текстов.

Отметим, что данная система обеспечивает возможность извлечения знаний из текстовой информации, которая представляется в виде соответствующих предикатов словосочетаний и правил продукций.

Первым уровнем обработки после выделения данных элементов текста является морфологический анализ. Он производится с использованием специальных морфологических словарей, которые содержат все словоформы данного языка с указанием их морфологических характеристик. На этой основе с использованием знаний экспертов (эвристик) формируется база знаний для выполнения морфологического анализа текстов на каждом из указанных языков. В результате стохастическому индексу каждого слова текста добавляется его лингвистический индекс, в который на данном этапе анализа заносятся его морфологические характеристики.

На втором уровне проводится синтаксический анализ, который реализуется с помощью специальной базы знаний, представленной в виде правил продукций, обеспечивающей синтаксический разбор простых и сложных предложений текста. При этом в лингвистический индекс каждого слова заносятся соответствующие синтаксические коды, определяющие данное слово как член предложения. Отметим, что параллельно с синтаксическим анализом членов предложения должен проводиться их семантический анализ, без которого невозможно определение членов предложения.

Семантический анализ текста начинается с автоматически выполняемой классификации общего словаря и специальных толковых словарей терминов и определений по заданным предметным областям, которые связаны с тематикой данной аналитической системы.

При классификации активно используются аналитические функции индуктивного и дедуктивного анализа связи слов, обрабатываемых в толковых словарях. В результате образуются семантические классификаторы, представленные в виде таблиц. Входом в таблицы являются стохастические индексы основ слов, строки таблицы содержат иерархию подклассов каждого слова и конечный класс, к которому данное слово принадлежит. Поскольку классификатор сделан для всех частей речи словарей, он позволяет определять типы, а также подклассы и классы объектов и субъектов предметной области, включая связи между ними.

С помощью классификатора формируются правила продукций для реализации параллельно синтаксического и семантического анализа текста, которые записываются в специальную базу знаний. После проведения пословного синтаксико-семантического анализа лингвистический индекс каждого слова дополняется его синтаксическими и семантическими характеристиками. В результате этого завершается процедура лингвистического анализа текста, после которого каждое слово каждого предложения будет представлено двумя стохастическими индексами: уникальным стохастическим индексом – идентификатором и лингвистическим индексом данного слова, содержащего все его морфологические, синтаксические и семантические характеристики, необходимые для дальнейшей индексации и разбора.

После этого переходят к автоматическому построению таблицы индексов данного текста в составе локальных, корпоративных баз данных или сайтов Интернет. Левый столбец таблицы содержит индексы неповторяющихся основ слов, входящих в текстовые документы по данной тематике, а строки содержат лингвистический индекс и адресную часть в виде совокупности индексов названия текстовых документов, индекса абзаца, предложения и предиката, в котором содержится данный индекс слова. Таблицы индексов текста используются при первичном поиске ответов или необходимых предложений текста с применением ключевых слов. Поиск по ключевым словам является основой для реализации второго уровня поиска с использованием семантики, извлечения знаний из текстов и аналитики.

Затем переходят к формированию концептуального описания предметной области текстов на основе выделенных в стохастической форме предикатов. Концептуальное описание представляется также в виде таблицы. Левый столбец содержит стохастические индексы всех неповторяющихся словосочетаний и предикатов индексируемого текста, строки включают индексы типов объектов и отношений между ними, а также (с использованием классификаторов) соответствующие им классы. Кроме этого, в состав таблицы также входит адресная часть, включающая индексы текста, абзаца и предложения, куда входят предикаты, которые содержат указанные классы объектов и отношений между ними. Это позволяет, используя классификатор и концептуальное описание предметной области, производить более точный повторный поиск необходимой информации после выполнения поиска по ключевым словам с тем, чтобы более полно и точно находить необходимые ответы или предложения, используя близкие по смыслу слова, словосочетания и предикаты, активно применяя семантический анализ текста.

На основе сформированного концептуального описания предметной области текста, а также используя формализованное описание функций определения, обобщения, сравнения, выбора, аналогии, дедукции и индукции, анализа и синтеза автоматически формируются правила продукций, содержащие необходимые типы и классы логически связанных предикатов предметной области текста. На основе этих функций могут формироваться деревья логического вывода, содержащие комбинации исходных логических функций, которые требуются пользователю системы для получения результата аналитического анализа с целью формирования обобщенных семантических характеристик словосочетаний, предикатов и сформированных из них предложений текста. Отметим, что предикаты, формируемые после выполненного лингвистического анализа, будут использоваться для эволюционного развития описания предметной области – «картины мира». Это обусловлено тем, что непосредственно к декларативной составляющей текстовых баз добавляются новые знания, извлекаемые из текста с помощью базовых аналитических функций и их заданных комбинаций. За счет комбинаций базовых функций исходная аналитическая система может автоматически настраиваться на заданную предметную область и эффективно использоваться в той области, к которой относится вводимая информация: например, управление, социальное обеспечение, финансирование, образование, культура, спорт и другие.

Для извлечения знаний из больших объемов неструктурированных текстов различных типов (диссертации, монографии, учебно-методическая, справочно-энциклопедическая литература и др.), аналитическая система может работать в автоматическом вопросно-ответном режиме. Здесь могут применяться разные варианты работы, например, осуществление точного семантического поиска, если информация непосредственно содержится в тексте и может быть выдана по запросу.

В более сложных случаях автоматически реализуются аналитические функции, которые после предварительной обработки информации с использованием процедур логического вывода, эквивалентных преобразований дают ответы на поставленные вопросы. Доказано [9], что если в системе может быть синтезирован алгоритм, который выдает ответ на поставленный вопрос с применением индексированной текстовой базы, то может быть создан аналитический алгоритм с использованием комбинаций разных функций, который обеспечит представление пользователю заданной информации.

В результате повышается эффективность формирования «картины мира» и обеспечивается полнота представленных понятий и связей между ними. На основе полученных предикатов, входящих в картину мира, автоматически формируются правила продукций по различным проблемным областям. В этом случае между предикатами семантической сети, которые отображают картину мира, выделяются семантические связи типа «условие-заключение», «причины-следствия», цели, определения и другие.

Как известно, правила продукций представляют собой символьную конструкцию вида «если (условие), то (заключение)». При этом условия содержат совокупность предикатов, объединенных логическими связками «и», а заключение содержит предикат, который выполняется, если все предикаты, входящие в условие, являются истинными для какой-то конкретной ситуации, соответствующей исследуемым объектам или процессам в определенной области знаний. Все полученные правила автоматически проверяются на их смысловую корректность. После стохастической индексации записываются в базы знаний.

Как было отмечено выше, представление словосочетаний, предикатов картины мира и правил продукций в стохастически индексированном виде дает возможность использовать эффективные алгоритмы логического вывода, а также (с помощью стохастической информационной технологии) исключить проблему комбинаторного взрыва. Без решения этой проблемы построение описанной выше системы лингвистического анализа текста и понимания смысла в принципе невозможно.

Методы понимания неструктурированной текстовой информации на основе полного лингвистического анализа

Как было показано выше, обработка текстов и знаний, входящих в картину мира, связана с решением проблемы BigData. Поэтому понимание смысла является важнейшей нерешенной задачей при создании автоматических (способных функционировать без участия человека) систем ввода и обработки текстовой, а также и сенсорной информации.

Отметим, что под пониманием смысла поступающих знаний и сенсорной информации подразумевается способность их интерпретации (представления) с использованием иных терминов той же самой знаковой системы или какой-либо другой (прежде всего языковой)[1].

Так, например, понимание смысла некоторого высказывания эквивалентно его переформулировке с использованием других терминов (иных слов) с полным сохранением смысла.

Создание полноценных систем понимания смысла текстов, речи и изображений невозможно без реализации функции автоматического самообучения при извлечении знаний из информационных сообщений и требует обеспечения возможности формирования нового знания, а также органичного (автоматического) дополнения этим знанием соответствующей опорной базы [2 - 6].

Представленные в предыдущем разделе методы полного лингвистического анализа, включая морфологический, синтаксический и семантический его уровни, обеспечивают получение смысла отдельных элементов предложений (словосочетаний и предикатов). Для перехода к анализу смысла предложений в целом, а также отдельных фрагментов текста, как указано в работе [1], его необходимо выполнять на прагматическом уровне (переход в область семиотики). В данном разделе предложен метод понимания смысла на прагматическом уровне с использованием описанной выше системы извлечения знаний из текстов и их логической обработки. Данный метод непосредственно связан с реферированием текстов.

Как известно, реферирование – это процесс анализа и переработки текста, выделения основных элементов его содержания с последующим изложением в устной форме (синтез речи) или в письменной форме (текстовое сообщение) [1]. В качестве подобных элементов будем рассматривать знания, выделенные из текста после его полного лингвистического анализа. Прежде отношения типа «род-вид», «часть-целое», «причина-следствие», «условие-заключение» и др. Далее используются аналитические функции определения, обобщения, сравнения, выбора, аналогии, дедукции и индукции, анализа и синтеза [2, 12]. В результате будут автоматически сформированы знания в виде правил продукций, элементов семантических сетей «картины мира», содержащие необходимые типы и классы предикатов предметной области текста. На основе этих знаний могут формироваться деревья логического вывода, включающие заданные комбинации исходных логических функций. Таким образом, производится первый этап реферирования текста на базе выделения из него существенно значимых элементов в виде знаний. Эти знания занимают значительно меньший объем, чем исходный реферируемый текст. Они позволяют получить его содержание в виде семантически связанных деревьев логического вывода [2, 12].

Следовательно, одной из основных проблем, возникающих при понимании смысла и реферировании, является извлечение знаний из текста, представление его в виде правил продукций и определение возможных траекторий логического вывода на множестве правил продукций. В существующих системах неструктурированного текста это связано с проблемой BigData, которая не подвластна современным информационным

технологиям. Вместе с тем, как было доказано в работах [2 – 4, 11], эта проблема успешно решается с помощью стохастики.

Метод понимания и реферирования текста на основе извлечения и обработки знаний

Для решения данной проблемы в [2] предложен метод определения возможных траекторий, поиска целей и предварительного выбора пути логического вывода, основанный на построении сети правил продукций и оперативного режима их обработки с использованием стохастических индексов и кодов. За счет применения стохастической информационной технологии, как будет доказано ниже, проблема комбинаторного взрыва полностью снимается, поскольку время логического вывода будет линейно зависеть от числа правил продукций, непосредственно задействованных в обработке. Отметим, что наполнения базы знаний правилами продукций может производиться в автоматическом режиме с использованием самообучающихся интеллектуальных систем извлечения знаний из текста, описанных в работах [3, 4].

Для реализации указанной возможности эффективной обработки знаний, извлекаемых из текста при его реферировании, каждое правило продукций подвергается стохастическому преобразованию и имеет следующий вид:

(1)

где - стохастические индексы предикатов условия,

- индекс предиката заключения правила.

Если предикат терминальный, то он связан посредством индекса с атомарной формулой вида , где - соответственно коды атрибутов u_i и их значений a_i фреймов или реляционных отношений, θ - арифметический оператор сравнения ($\geq$, $\leq$, $=$ и др.).

В процессе обработки терминального предиката , по индексу производится доступ к соответствующему фрейму или реляционному отношению опорной базы знаний. При этом истинность или ложность предиката устанавливается путем сравнения кодов атомарной формулы, определяемой индексом с кодом фрейма или реляционного отношения. Сравнение индексов и кодов осуществляется с использованием стохастического процессора, обеспечивающего выполнение логических и арифметических функций со стохастическими индексами и кодами без их расшифровки [9]. Реализация процедур обработки терминальных предикатов путем произвольного доступа по стохастическим индексам к соответствующим фреймам и реляционным отношениям, а также выполнение функций сравнения стохастических индексов и кодов в стохастическом процессоре обеспечивает высокую эффективность обработки терминальных предикатов. Это относится и к обработке словосочетаний и предикатов

данной предметной области. Указанные параметры и значения после стохастической индексации записываются в соответствующие фреймы базы знаний.

Проблема устранения комбинаторного взрыва решается, как было отмечено выше, на основе автоматического формирования метазнаний системы логического вывода в режиме самообучения. Здесь применяются агрегативные свойств стохастических индексов. В данном случае стохастический индекс каждого правила продукций вычисляется путем сложения по mod 2 индексов всех m предикатов каждого правила:

(2)

Таким образом, стохастический индекс правила связан со стохастическими индексами составляющих его предикатов как «целое-часть». На этой основе, для построения сети правил продукций системы логического вывода, автоматически формируются метазнания - специальные сетевые фреймы, в которые включаются стохастические индексы правил, имеющих идентичные предикаты в условии или заключении.

Указанные сетевые фреймы формируются для каждого правила базы знаний. При этом индексы правил , имеющих предикат в условии или заключении, семантически идентичный предикату , будут включены в один и тот же сетевой фрейм. Это обусловлено тем, что доступ к сетевым фреймам осуществляется по индексам предикатов . Поэтому все индексы правил , относящиеся к одному и тому же предикату, будут содержаться в одном и том же сетевом фрейме. Объединенные логическими связками ИЛИ они определяют альтернативные направления логического вывода.

Образованная сетевая структура позволяет реализовать прямую и обратную последовательности логического вывода, используя только связанные друг с другом семантически идентичными предикатами правила продукций. Тем самым исключается необходимость перебора на всем множестве правил продукций при выполнении каждого очередного цикла логического вывода. В результате исключается главная причина комбинаторного взрыва. При этом за счет использования сетевых фреймов при построении траектории логического вывода каждое правило обрабатывается не более одного раза. При реализации любой возможной траектории логического вывода на множестве правил продукций может быть задействовано не более общего числа M<<N правил продукций в базе знаний. Поэтому время логического вывода с использованием предложенного метода на основе стохастической информационной технологии будет линейно зависеть от числа правил продукций, непосредственно входящих в траекторию логического вывода в базе знаний:

(3)

где - время логического вывода на множестве правил продукций,

а - время обработки правила продукций.

Покажем, что применение сетевых правил продукций позволяет формировать с помощью логического вывода текстовые сообщения,

эквивалентные полученному исходному текстовому сообщению для понимания его смысла. Все эти сообщения соответствуют тексту, полученному в процессе распознавания смысла с использованием семантического классификатора. Например, для исходного текстового сообщения «футболист бежит по полю» и полученного обобщения с использованием семантической классификации картины мира «человек перемещается в пространстве» с помощью логического вывода может быть получено определенное число эквивалентных сообщений. Таким образом, описанная стохастическая система на основе логического вывода полностью реализует функцию понимания смысла информации в контексте определения [1].

Для эффективной реализации логического вывода в работе [2] предложен оперативный режим генерации дерева траекторий логического вывода. Он основан на применении процедур доступа по стохастическим индексам к сетевым фреймам и фреймам правил с отложенной обработкой терминальных предикатов (высказываний), требующих обращения к внешней памяти.

При реализации этого режима после выбора целевого правила , описывающего, например, обобщенное текстовое сообщение на каждом последующем уровне j дерева траекторий из сетевых фреймов образуется логическое выражение, включающее индексы правил, заключения которых связаны идентичными предикатами с условиями правил (j-1)-го уровня. Каждому правилу (j-1)-го уровня в общем случае может соответствовать набор сетевых фреймов, содержащих индексы правил, заключения которых связаны с условием правила . При формировании логического выражения уровня индексы правил внутри сетевых фреймов, объединенные логическими связками ИЛИ, заключаются в скобки. Сетевые фреймы, относящиеся к одному правилу (j)-го уровня, закрываются дополнительными скобками. В результате будет сформировано логическое выражение вида:

, (4)

которое поступает в дальнейшую обработку.

Как показано в работе [2], применение режима оперативной генерации траекторий логического вывода с помощью сетевых фреймов дает возможность формирования исходного текстового сообщения и его возможных текстовых эквивалентов в виде полученных составных продукций, обеспечивающих достижение цели логического вывода. Проверка истинности полученных эквивалентов и выбор из них предложений, соответствующих моделируемой ситуации, производится в процессе обработки их терминальных предикатов. При этом, поскольку выделенные цепочки правил продукций содержат только основные элементы содержания

исходного текста, то одновременно в процессе логического вывода осуществляется и автоматическое реферирование исходного текста. В представленном ниже примере данный процесс будет подробно проанализирован.

Рассмотрим порядок реализации логического вывода и построения необходимых цепочек правил продукций с использованием базы знаний. В результате описанного выше алгоритма логического вывода с использованием сетевых правил образуется некоторое дерево логического вывода (Рис. 1). Допустим, что указанное дерево содержит три уровня логически связанных через сетевые фреймы правил продукций -. После генерации логического выражения на каждом уровне можем получить следующий результат:

(5)

Сформированное логическое выражение поступает в дальнейшую обработку для определения истинности предиката заключения целевого правила продукций.

Покажем, что применение режима оперативной генерации траекторий логического вывода с помощью сетевых фреймов, позволяет предварительно сформировать все необходимые цепочки логического вывода и определить исходное текстовое сообщение и все его семантические эквиваленты. При этом под исходным сообщением будем понимать одну из возможных траекторий (цепочек правил), обеспечивающую достижение цели логического вывода и содержащую терминальные предикаты, соответствующие предикатам полученного при распознавании текста сообщения.

Для этого используем понятие составной продукции[7], которая представляет собой совокупность всех терминальных предикатов цепочки правил продукций дерева траекторий логического

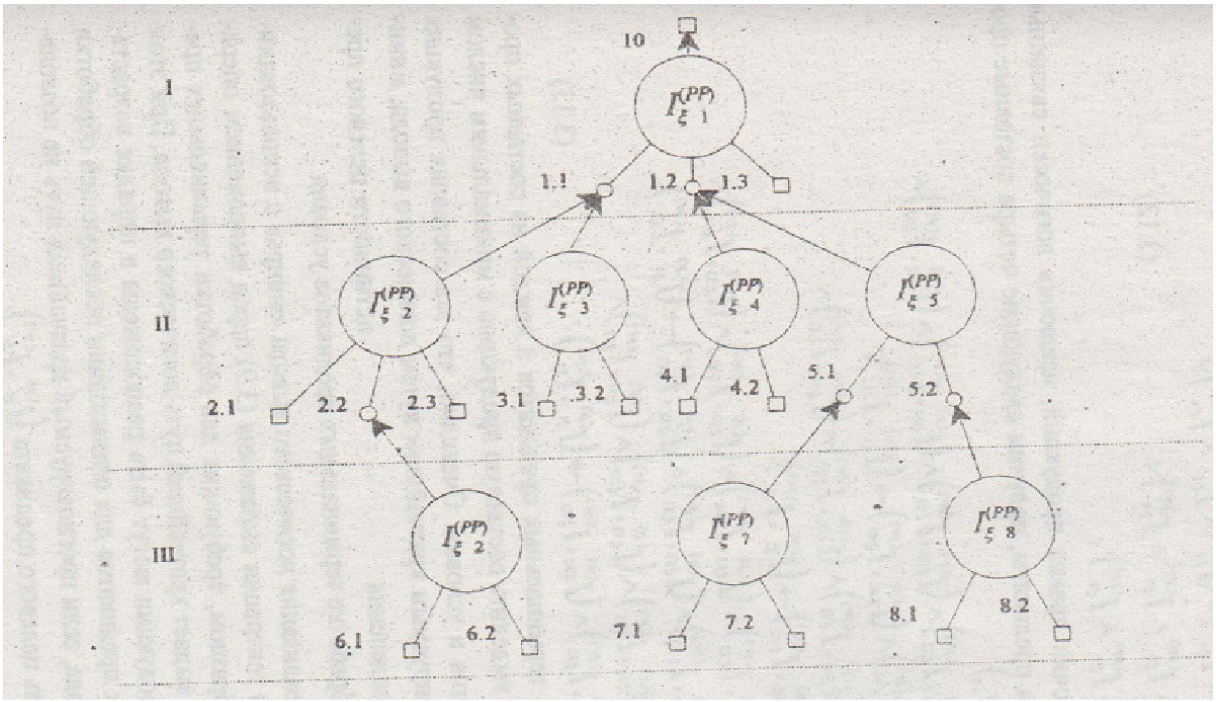

Рис. 1 Дерево логического вывода

вывода, объединенных логическими связками $\wedge$ и обеспечивающих независимое определение истинности или ложности целевого предиката . В соответствии с этим определением составная продукция имеет следующий вид:

, (6)

где - терминальные предикаты цепочки правил продукций;

- предикат заключения целевого правила продукций, определяющий имя обобщенного предиката дерева логического вывода.

Легко видеть, что совокупность составных продукций может быть получена из выражения (4), сформированного в результате оперативной генерации дерева траекторий логического вывода после раскрытия скобок.

Для рассматриваемого примера из сформированного выражения можно получить четыре независимые цепочки логического вывода:

(7)

Из каждой цепочки правил продукций, исключив логически связанные через сетевые фреймы предикаты, получим следующие четыре составные продукции (P_{s1} , P_{s2} , P_{s3} и P_{s4}):

(8)

Сравнивая терминальные предикаты в каждой из составных продукций Psi, мы можем выделить составную продукцию с терминальными предикатами исходного текстового сообщения. Например, будем считать, что продукция Ps4 соответствует этому условию.

Выделение из текста знаний в виде правил продукций как основных содержательных элементов, и представление их как цепочек дерева логического вывода реализует, по существу, первый уровень реферирования заданного текста. При этом получение составных продукций позволяет представить содержание реферируемого текста в более обобщенном виде.

Таким образом, реализация логического вывода "вширь" с использованием режима оперативной генерации выражения (4) позволяет сформировать реферат исходного текстового сообщения. Остальные составные продукций {Ps2i}могут быть выделены в качестве семантических эквивалентов данного реферата после обработки терминальных предикатов базы знаний, требующих обращения к внешней памяти. Следовательно, изложенный метод, разработанный на основе метода логического вывода [2], обеспечивает понимание смысла текстовых сообщений с использованием картины мира, системы семантической классификации и базы знаний. Одновременно логическая обработка автоматически извлекаемых из текста правил продукций позволяет реализовать автоматическое реферирование исходного текста. Этот процесс более подробно можно изучить на основе представленного ниже примера.

Пример реферирования текстового сообщения на основе логического вывода

Рассмотрим пример реферирования отрывка текста из викторины «Города и реки России»: «В речном заливе было много яхт под парусами. По-видимому, здесь проходила парусная регата. Это свидетельствовало о том, что в городе развит парусный спорт. По мере движения к центру выясняется, что в городе есть речные вокзалы и мосты, которые разводятся. Из этого следует, что через город по реке проходят крупные суда. Поскольку в городе развит парусный спорт и через город проходят крупные суда, то, очевидно, что город стоит на крупной реке. Из путеводителя следует, что в городе проживает несколько миллионов человек. Следовательно, река протекает через крупный город. Из сказанного можно сделать вывод, что крупный город стоит на крупной реке и этот город (ответ) – Санкт-Петербург, а протекающая через город река, - Нева».

После извлечения знаний из данного фрагмента текста в соответствии с описанным выше методом (при этом используется картина мира, система семантической классификации) получим следующий набор правил, описывающих возможные варианты реферирования данного текста.

Состав выделенных из текста знаний в виде правил продукций:

1. Если через город протекает крупная река,

и город расположен на крупной реке,
и в городе большое население,
то крупный город стоит на большой реке.

2. Если в городе есть речное сообщение,
и в городе есть мосты, соединяющие части города,
то через город протекает большая река.
3. Если в городе есть мосты,
и эти мосты разводные,
то через город протекает большая река.
4. Если река делит город на несколько частей,
то город расположен на реке.
5. Если в городе развит речной спортивный флот,
и через город проходят большие суда,
то город расположен на крупной реке.
6. Если в городе есть причалы и речные трамвайчики,
то в городе есть речное сообщение.
7. Если в городе есть речные вокзалы,
и мосты в городе разводятся,
то через город проходят крупные суда.
8. Если в городе развит парусный спорт и проводятся регаты,
то в городе развит речной спортивный флот.

На основе данного множества правил продукций, используя описанный выше метод, получим четыре составные продукции следующего содержания:
1. Если в городе есть причалы и речные трамвайчики,
а также мосты, которые разводятся,
и город имеет большое население,
то крупный город стоит на большой реке.
2. Если в городе есть мосты,
и эти мосты разводные,
и город расположен на крупной реке,
и в городе большое население,
то крупный город стоит на большой реке.
3. Если река делит город на несколько частей,
и город расположен на большой реке,
и в городе большое население,
то крупный город стоит на большой реке.
4. Если в городе есть речные вокзалы,
и мосты в городе разводятся,
и в городе развит парусный спорт и проводятся регаты,
и город расположен на крупной реке,
и в городе большое население
то крупный город стоит на большой реке.

Отметим, что четвертая составная продукция соответствует исходному текстовому сообщению. Остальные составные продукции являются

возможными эквивалентами исходного сообщения. Для проверки этой гипотезы необходимо провести, как было указано выше, обработку их терминальных предикатов с использованием базы знаний. При этом составные продукции, содержащие только истинные терминальные предикаты, являются семантическими эквивалентами исходного текста в процессе реферирования.

Таким образом, данный метод реализует функцию реферирования на основе извлечения знаний из текста, что позволяет перейти к автоматическому семантическому реферированию неструктурированного текста с использованием метазнаний.

Метод реферирования текста на основе формирования функциональных метазнаний

Для реализации функций реферирования текста на больших пространствах поиска правил продукций и картины мира в систему вводится дополнительный уровень функциональных метазнаний, содержащий метаправила и метафакты. Метаправила образуются автоматически, путем агрегации фрагментов сетевой структуры правил продукции, ограниченных сетевыми фреймами, определяющими альтернативные направления логического вывода, или терминальными предикатами.

При синтезе метаправил от каждого целевого правила или от правила, заключение которого связано с сетевым фреймом, определяющим альтернативные направления вывода, производится генерация составных продукций , имеющих в качестве заключения предикат правила . Полученное метаправило имеет вид:

(9)

За счет введения метаправил-агрегатов и соответствующих сетевых фреймов сеть правил продукций заменяется сетью метаправил. Создание уровня метаправил-агрегатов позволяет перейти к формированию метафактов, являющихся результатами обработки терминальных предикатов каждого метаправила с использованием фактуальных знаний, которые в соответствии с концептуальным описанием базы знаний относятся к данному метаправилу. С этой целью, для каждого высказывания , сформированного из предиката заключения метаправила производится генерация дерева логического вывода в рамках данного фрагмента метаправила и обработка составных продукций с использованием фактуальных знаний. В результате, фрейм метафактов каждого метаправила будет содержать значения составных продукции для каждого конкретного высказывания , образованного из предиката заключения данного метаправила.

Введение уровня метаправил и метафактов позволяет создавать логические структуры, предназначенные для семантического реферирования неструктурированного текста и гипотез распознаваемых сообщений для разговорной русской речи [2, 11, 12].

В соответствии с этим каждая составная продукция характеризуется выбором семантически значимых элементов текста, которые необходимы для краткого его изложения. Таким образом, мы переходим на уровень семиотического анализа текста, где в качестве значимых элементов используются типовые состояния, ситуации, действия в разных проблемных областях и другие типовые семантические структуры. Логический вывод на уровне метаправил обеспечивает принципиально новые возможности автоматического реферирования объемных текстов.

Важным свойством предлагаемого метода автоматического реферирования текста является возможность «усиления» функции обобщения за счет использования дополнительных семиотических структур. Это происходит, когда первоначально выбранное метаправило не позволяет достаточно полно изложить основную сущность текстовых сообщений, полученных после обработки текста или распознавания речи. В этом случае образуется фрагмент сети, включающий несколько метаправил. Они связаны с одним и тем же целевым предикатом, который определяет семантическое обобщение данного фрагмента текста.

Все составные продукции, входящие в эти метаправила, являются независимыми друг от друга и могут использовать различные семантические элементы или методы для изложения сущности поступившего сообщения. В общем случае дополнительные продукции могут включать эквивалентные преобразования исходного текста.

На Рис. 2 представлен пример образования метаправил на основе сетевой структуры правил продукций. При этом на Рис. 2а показана сформированная сетевая структура, полученная в процессе реферирования текстового сообщения, а на Рис. 2б – результат агрегации данной структуры в виде метаправил, включающих терминальные и целевые предикаты, отображающие содержание полученного реферата.

Отметим также, что возможность реализации логического вывода на метауровне с использованием метаправил и метафактов позволяет существенно повысить оперативность обработки знаний и данных за счет значительного сокращения числа обращений к внешней памяти при выполнении процедур сопоставления с образцом.

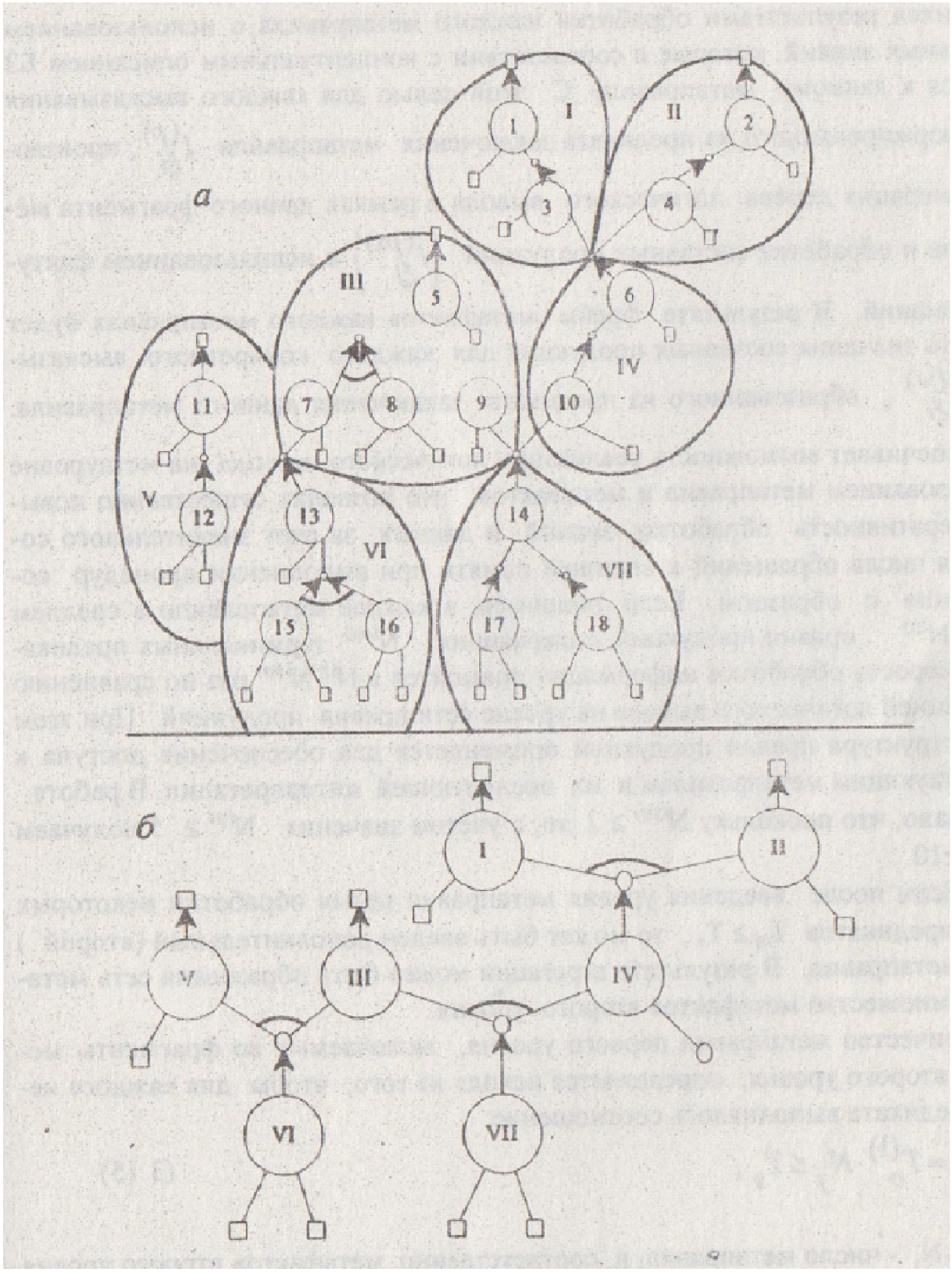

Рис. 2. Синтез метаправил:

а – сетевая структура, включающая фрагменты метаправил (I-VII);
б – структура сети метаправил-агрегатов

Если, например, в каждое метаправило в среднем входит $N^{(p)}$ правил продукций, содержащих $N^{(pp)}$ терминальных предикатов, то скорость обработки информации повысится в $N^{(p)}N^{(pp)}$ раз по сравнению с реализацией логического вывода на уровне сети правил продукций. При этом сетевая структура правил продукций применяется для обеспечения доступа к соответствующим метаправилам и их последующей интерпретации. В работе [2] показано, что поскольку $N^{(pp)} \geq 2$, то, с учетом значения $N^{(p)} \geq 5$ получаем $N^{(p)}N^{(pp)} \geq 10$.

Если после введения уровня метаправил полученное текстовое сообщение не является достаточно кратким и не может быть обработано в заданное время $T_{oj} \geq T_з$, то может быть введен дополнительный (второй) уровень метаправил. В результате агрегации может быть образована сеть метаправил и множество метафактов второго уровня.

Количество метаправил первого уровня, включаемых во фрагменты метаправил второго уровня, определяется исходя из того, чтобы для каждого целевого предиката выполнялось соотношение:

, (10)

где N_j − число метаправил, и, соответственно, метафактов второго уровня в дереве логического вывода, сформированного от данного целевого предиката, − время обработки одного метафакта, определяемого временем доступа к записи соответствующего метафакта. Таким образом, на основе ввода двухуровневой структуры функциональных метаправил может быть обеспечено $T_з$ заданное время обработки полученных текстовых сообщений путем их эффективного реферирования.

Важным свойством образованной двухуровневой сетевой структуры подсистемы логического вывода является возможность ее автоматического оперативного реструктурирования за счет модификации сетевых фреймов при вводе новых или исключении правил продукций и метаправил. Это обеспечивает универсальность предложенных методов автоматического реферирования текстов для любой предметной области, включая тексты больших объемов в режиме BigData. Представленный ниже пример иллюстрирует процесс автоматического реферирования текстов с использованием метазнаний.

Рассмотрим пример реферирования текста, полученного в процессе распознавания речи комментатора футбольного матча.

Содержание фрагмента:

1. За десять минут до окончания матча тренер нашей сборной произвел замену двух нападающих на защитника и полузащитника. После этого вся команда, за исключением центрального нападающего переместилась на свою половину поля и организовала две линии защиты, перекрыв все подходы к своим воротам. Таким образом, сборная перешла к явно выраженной обороне своих ворот.

Поскольку счет в данном матче был «ничейный» (1:1), а в предыдущем матче отборочного тура сборная одержала победу, то полученный результат обеспечивал выход сборной в следующий тур соревнований. Если матч

закончится с указанным «ничейным» счетом, то, по мнению тренера, стоит поберечь силы команды для успешной игры в следующем туре соревнований.

В результате применения методов логического вывода с использованием знаний, извлеченных из данного речевого сообщения после его распознавания, и сформированных метаправил может быть автоматически получен следующий реферат данного текста:

2. В конце матча наша сборная перешла к обороне. «Ничейный» результат матча позволял команде выйти в следующий тур отборочных соревнований. Тренер поберёг силы команды.

Полученный реферат сформирован автоматически, содержит основные положения исходного текста репортажа и обеспечивает четырехкратное сокращение объема текста.

Таким образом, на основе стохастики обеспечиваются функции понимания смысла и извлечения знаний из текста как основы для автоматического реферирования текстовой информации в различных проблемных областях. Это позволяет достичь универсальности разработанного метода автоматического реферирования, что обеспечивает мировой уровень новизны предложенных методических и технологических решений. Как следует из описания процесса реализации указанных функций, он может быть осуществлен с использованием квантового компьютера с интеллектом, представленного в разделе 4. Сделанный вывод открывает принципиально новые перспективы в развитии информатики и индустрии знаний в 21 веке.

Литература

1. Потапов А.С. Распознавание образов и машинное восприятие. Спб.: Политехника, 2007. - 548с.
2. Насыпный В.В. Развитие теории построения открытых систем на основе информационной технологии искусственного интеллекта. М.: Воениздат, 1994. - 248с.
3. Насыпный В.В., Насыпная Г.А. Способ синтеза самообучающейся системы извлечения знаний из текстовых документов для поисковых систем. Патент РФ №2273879, номер международной заявки PCT/RU02/00258, дата подачи 28 мая 2002.
4. Насыпный В.В., Насыпная Г.А. Способ синтеза самообучающейся аналитической вопросно-ответной системы с извлечением знаний из текстов, заявка на патент №2007120344/09 от 06.08.2007. Получено решение на выдачу патента на изобретение от 21.07.2008.
5. Современный русский язык: Учеб.дляфилол. спец. высших учебных заведений. Под редакцией В.А. Белошапковой. М.: Азбуковник, 1999. – 928с.
6. Насыпный В.В., Насыпная Г.А. Система распознавания, понимания смысла, анимационного моделирования и синтеза речи на основе стохастической информационной технологии. М.: Прометей, 2008. – 76 с.

7. Искусственный интеллект. Справочник. Кн. 2. Модели и методы. Под ред. Поспелова Д.А. М.: Радио и связь, 1990. - 303 с.
8. Halsall F. Data communications computer networks and osi. Addison-wesley publishing company, 1988. - 973 с.
9. Насыпный В.В. Способ комплексной защиты распределенной обработки информации в компьютерных системах и система для осуществления способа. Патент РФ №2259639, номер международной заявки PCT/RU /00272, дата подачи 28.10.2003г.
10. Насыпный В.В., Насыпная Г.А. Метод семантической связи текста с трехмерной графикой. – М.: Прометей, 2007. – 27 с.
11. Насыпный В.В. Распознавание и понимание смысла речи в шумах на основе стохастики. - М.: МПГУ, Прометей, 2012. – 129 с.
12. Насыпный В.В. Стохастика как основа для перехода к большим данным, индустрии знаний и нанотехнологии. – М.: МПГУ, 2011. – 24 с.

6. Распознавание речи и видеоинформации

Речь

В настоящее время активно проводятся исследования по созданию системы и технологий распознавания речи, которые могли бы эффективно преобразовывать вводимую в компьютер слитную речь от неизвестного диктора в корректное текстовое сообщение с гарантированным уровнем достоверности в реальном масштабе времени. При этом важнейшим требованием является реализация этих процессов в условиях шумового воздействия на канал связи, естественной вариативности темпа и громкости речи, а также амплитудно-частотных искажений в канале ее приема-передачи и т.п. Необходимо также обеспечить достоверность распознавания речи при наличии у неизвестного диктора иностранного акцента, местного диалекта, индивидуальных особенностей речеобразования.

Выполненный анализ существующих систем и технологий обоснованно показывает невозможность использования для достоверного распознавания речи современных статистических методов и прежде всего скрытых марковских моделей в условиях нормализации сигнала и реализации функций помехозащищенности.

Особенно важной является разработка систем, предназначенных для выделения в потоке слитной речи определенного состава слов и словосочетаний по заданной тематике в условиях шума. Этот состав ключевых слов может являться перечнем команд административного или производственного управления. После перевода в текстовое сообщение ключевые элементы обрабатываются в компьютерах в автоматическом или автоматизированном режиме. Очевидно, что к

таким системам предъявляются высокие требования по достоверности распознавания речи в шумах.

Однако применяемые в существующих системах методологии и технологии распознавания речи не дают ощутимых результатов, достаточных для создания систем государственного назначения или коммерческого применения.

Поэтому, на наш взгляд, для решения этой сложнейшей научно-технической проблемы нужны принципиально новые подходы и технологии. Они должны быть направлены, прежде всего, на моделирование тех процессов, которые осуществляет человек при речевом общении в условиях шумовых воздействий.

Основным выводом из анализа современного состояния рассматриваемой проблемы, на наш взгляд, является то, что ее невозможно решить автономно и без выполнения функций нормализации сигнала. Эта проблема может быть успешно решена только в едином комплексе распознавания, нормализации, понимания смысла и синтеза речи с использованием нового поколения самообучающихся интеллектуальных систем извлечения знаний из текстовой информации и речевых образов. Отметим, что создание эффективных интеллектуальных систем с применением традиционной информационной технологии является в настоящее время практически неразрешимой задачей из-за «комбинаторного взрыва», который возникает вследствие переборного механизма логического вывода [1].

В основу эффективных интеллектуальных систем, отвечающих необходимым требованиям, могут быть положены способы и технологии, описанные в работах [2, 3, 4, 6, 9]. В этих работах показано, что создание отмеченных интеллектуальных систем, обеспечивающих возможность логической обработки больших объемов знаний текста и речевых образов в реальном времени и их нормализация, возможны на базе применения отечественной стохастической информационной технологии. Использование [2] данной технологии позволяет исключить «комбинаторный взрыв» при реализации индуктивного логического вывода на значительных пространствах символьной информации и обеспечить линейную зависимость времени логической обработки от числа элементов знаний, задействованных в логическом выводе. Это свойство в сочетании с механизмами самообучения позволяет автоматически создавать и эффективно использовать в процессе распознавания речи, понимания ее смысла и синтеза речевых сообщений большие базы знаний, которые на новом качественном уровне решают указанные сложнейшие проблемы.

Кроме этого, как показано в работах [2, 9], стохастическая информационная технология в силу своих свойств обеспечивает высокую

эффективность распознавания речи в шумах и при вредоносном информационном воздействии на систему (атаки хакеров, компьютерные вирусы и закладки). Это позволяет создавать принципиально новые помехоустойчивые интеллектуальные системы распознавания речи.

Во главу угла при создании методов распознавания речи положено использование многоуровневых интеллектуальных систем. Они обеспечивают эффективную акустическую и артикуляционную классификацию вводимой речевой информации, выделение в ней различных типов звуков и слогов, определение границ слов, а также вычленение предложений из непрерывной речевой информации при шумовом воздействии.

В соответствии с предложенным образно-семантическим методом[6] в процессе распознавания речи создается система опорных и классификационных семантических кодов, которые взаимооднозначно определяют звуковые образы независимо от диктора. Эти коды синтезируются также и для ключевых слов, что позволяет выделять их в потоке слитной речи с заданной достоверностью.

Многоуровневые интеллектуальные системы обработки информации (снизу вверх) с аппаратом логического вывода на основе баз знаний позволяют описывать и извлекать соответствующие фонетические и текстовые структуры из речевых сигналов, используя при этом фонетический, лексический, морфологический, синтаксический, семантический и прагматический виды анализа.

Одновременно с этим (сверху вниз) проводится синтез звуковых сигналов и речевых сообщений, которые непосредственно связаны с текстовыми вариантами распознавания акустического сигнала. Для этого производится генерация речевых образов, базирующихся на текстовых сообщениях, которые близки по смыслу полученным вариантам текстовых структур при анализе речи снизу вверх.

Далее осуществляется коррекция синтезированных речевых сигналов для их максимального совпадения с полученным входным речевым сообщением. Здесь применяется программно-визуальная анимационная модель речевого тракта, которая обладает способностью адаптироваться к анатомическим особенностям органов речеобразования любого из дикторов.

Данная модель является одним из основных элементов системы синтеза речи. Она базируется на разработанном в [10] методе семантической связи текста с трехмерной графикой. Указанная модель, также используя соответствующие базы знаний и логический вывод, визуализирует артикуляционный процесс в тракте речеобразования с синхронной генерацией необходимого звукового образа. При этом реализуется основная функция коррекции трехмерного изображения типового речевого тракта с целью его адаптации под особенности речевых органов и артикуляционных параметров каждого конкретного диктора.

Здесь получило воплощение одно из важных положений науки о распознавании речи – теория внутренней модели, которая объединяет

процессы речеобразования и восприятия речи. Внутренняя модель, формируемая в сознании человека, может использоваться при распознавании речи других людей, дополняя пространство акустических признаков пространством артикуляторных параметров [11, 12].

Для выполнения встречного процесса анализа с использованием сближающихся по смыслу текстов и выделенных параметров речевых сигналов, которые получены при анализе и синтезе речи, самообучающаяся интеллектуальная система осуществляет глубокий семантический анализ результата обработки и синтеза речевых сообщений. С этой целью автоматически формируется и применяется «картина мира», содержащая словосочетания и предикаты по различным предметным областям с указанием их семантических классов. В результате удается резко повысить эффективность встречного, основанного на анализе и синтезе, процесса распознавания и понимания речи и обеспечить высокую достоверность распознавания текстов, соответствующих непрерывному речевому сигналу [14].

Подчеркнем, что при этом существенную роль играют стохастические коды, корректирующие ошибки [2], которые используются также для обработки речевых сигналов путем дополнительной адаптации синтезируемого сигнала под звуковые образы речевой информации, поступающей от данного диктора.

Таким образом, основой указанной концепции создания единого контура распознавания, нормализации, понимания смысла и синтеза речевых сообщений является стохастическая информационная технология. Именно за счет этой новой отечественной технологии достигается принципиально новая возможность интеллектуальной обработки речевых образов, их эффективное распознавание и коррекция с использованием встречного процесса акустического анализа и синтеза речи на основе непрерывно формируемого и уточняемого смыслового содержания поступающих в систему речевых сообщений и выделения ключевых элементов по заданной тематике.

Комплексное распознавание речи и видеоинформации

Как показано в работе [14], применение разработанных методов и технологий образно-семантической и семантико-параметрической обработки информации позволяет комплексно решать проблему распознавания речи и видеоинформации. При этом, как будет показано ниже, на уровне фонетического анализа речи возможно применение уже существующих систем распознавания образов для реализации некоторых функций распознавания речи. В данном разделе рассмотрена возможность использования нейросети Numenta не только для распознавания образов в соответствии с ее предназначением, но и для реализации некоторых функций распознавания звуковых образов в слитной речи. Для этого была использована схема построения комплекса распознавания речи, представленная на Рис.1. Эта схема позволяет реализовывать два контура

распознавания речи на уровне ее фонетического анализа. Первый контур предназначен для интеллектуального сканирования звуковых образов слитной речи от неизвестного диктора, представленного в виде осциллограммы и спектрограммы сигнала. Интеллектуальное сканирование основано на использовании семантико-параметрического метода распознавания речи и позволяет с помощью баз знаний, содержащих артикуляционное и акустическое описание звуковых образов, а также, используя логический вывод, проводить многоуровневую классификацию звуков, от семантики опорных кодов до распознавания фонем и собственно наименований звуков. С этой целью создан специальный классификационный словарь русского языка объемом около миллиона слов. В результате работы этого контура распознавания речи обеспечивается заданная достоверность выделения наименований звуков, слогов и отдельных слов в потоке слитной речи. С целью повышения достоверности распознавания речи до уровня практически 100% для любого произвольного диктора вводится второй контур, основанный на реализации образно-семантического распознавания речи с использованием системы распознавания образов, в данном случае нейросети Numenta.

На Рис.1 представлена структурная схема, позволяющая реализовать на фонетическом уровне два описанных контура анализа. Она включает несколько систем.

Первая из этих систем, подобная существующей Adobe Audition, предназначена для начальной обработки речевого сигнала в спектральной форме или в виде осциллограммы, в частности, для измерения различных параметров звуковых образов. К ним относятся прежде всего линейные размеры абрисов звуковых сигналов и их последовательностей, частотные и временные характеристики формант звуков, артикуляционных расстояний между звуками, форма огибающей осциллограмм и спектральной составляющей сигнала, взаимное расположение формант (компактное и диффузное) и другие параметры, которые необходимы для логической обработки речевых сигналов. Отметим, что именно в ходе логической обработки формируются первичные семантические образы звуковых сигналов, которые предварительно позволяют определить значение звуков и слов, входящих в слитную речь.

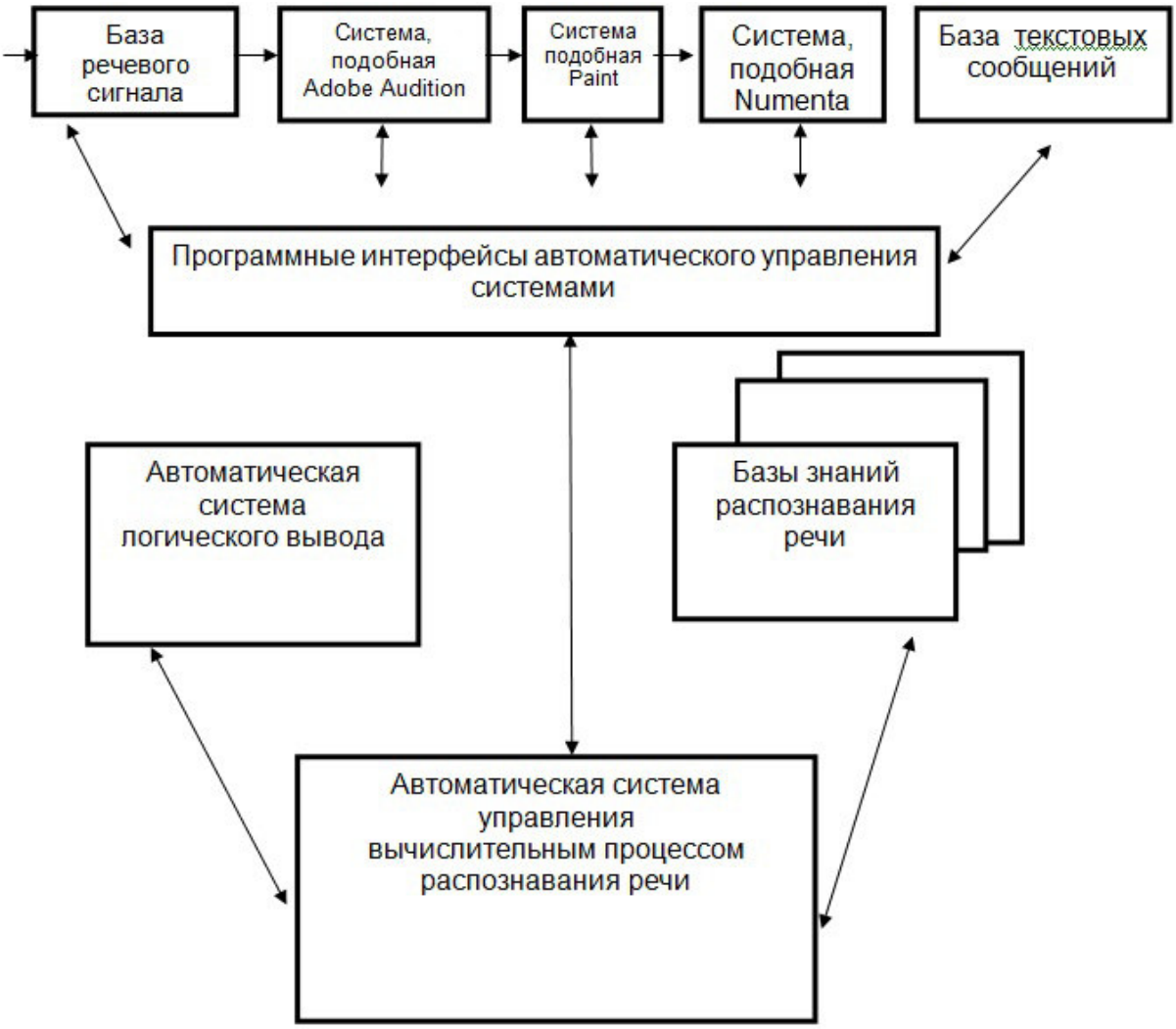

Рис. 1. Структура автоматического программного макета распознавания речи

Вторая система, подобная существующей Paint, предназначена для формирования и обработки рисунков, содержащих абрисы звуковых образов в спектрограмме или осциллограмме, а также их семантически важных фрагментов. Третья система, используемая в макете, – это система, подобная нейросети Numenta. В данном макете ее механизмы распознавания образов используются для получения семантических кодов звуков при их артикуляционном или акустическом анализе, а также при распознавании отдельных элементов звуков и их классов. Отметим, что нейросеть Numenta построена на основе наиболее перспективной технологии распознавания образов, использующей иерархическую темпоральную память (HTM).

Результаты обработки речевой информации поступают в систему логического вывода. Здесь в автоматическом режиме полученная информация применяется для реализации двух описанных контуров распознавания речи от неизвестного диктора.

Отметим, что лингвистическая и семантическая обработка полученной на данном этапе текстовой информации в ходе распознавания речи будет реализована на втором этапе создания опытного образца распознавания слитной речи от неизвестного диктора [14]. В результате описанного алгоритма обработки на фонетическом уровне, как было указано выше, достигается высокая (около 100%) достоверность распознавания речи. Совместно с этим могут быть эффективно использованы возможности системы Numenta для комплексного распознавания слитной речи от неизвестного диктора одновременно с распознаванием образа говорящего человека. Это продемонстрировано на Рис. 2.

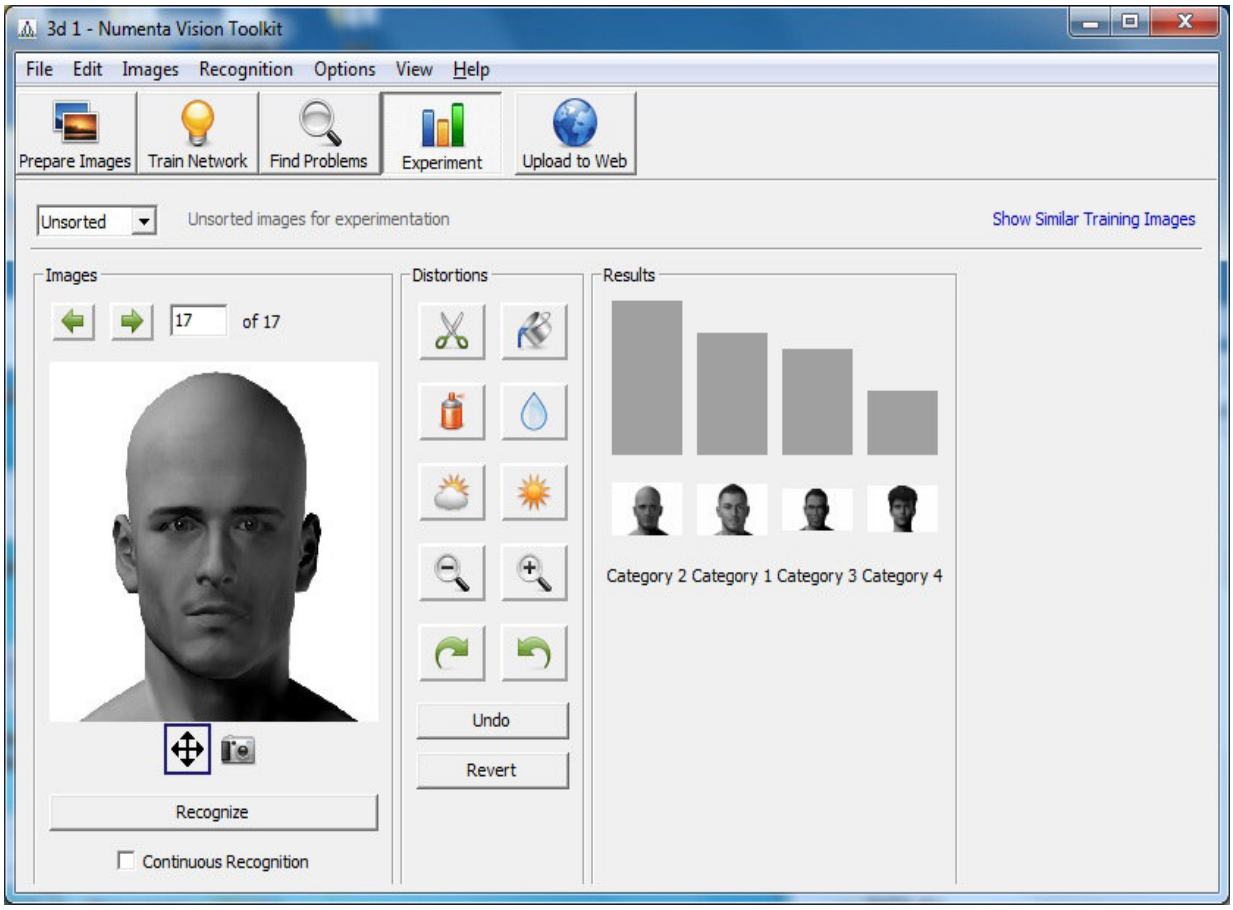

Рис. 2. Распознавание изображений лиц с помощью системы Numenta

Введенное изображение лица контролируемого пользователя уверенно распознается системой Numenta, о чем свидетельствует присвоение его распознанному образу Category 1 с наивысшим уровнем достоверности.

Далее показано использование системы Numenta для моделирования контроля произносимой речи данным диктором на основе образно-семантического метода. В ходе данного моделирования была подтверждена возможность применения системы Numenta для распознавания речи, включающей всю шкалы семантических кодов. Эти коды представлены в классификационном словаре, который подробно описан в работе [14]. В

результате было показано, что система Numenta позволяет уверенно распознавать все восемь уровней кодирования звуков слитной речи, реализуя при этом предложенный в работе [14] образно-семантический метод. Отметим, что до этого функционировал первый контур распознавания речи, который осуществлял процесс семантико-параметрического распознавания речи на основе интеллектуального сканирования звуковых образов. Поэтому при работе системы Numenta использовалась информация о значении предварительно распознанных кодов и звуков. В этом заключается совместное функционирование двух контуров распознавания речи – семантико-параметрического и образно-семантического.

Ниже представлен пример функционирования системы Numenta на восьмом уровне распознавания речи, а именно, при определении значений отдельных фонем и наименований звуков. С этой целью выбран класс гласных твердых звуков (А, О, И, У, Э, Ы), каждый из которых приведен в оригинальном физиономическом представлении, включающем наиболее характерные для данного звука элементы спектрограммы и осциллограммы. В результате формируется устойчивый, уникальный образ каждого звука, независящий от артикуляционных и акустических особенностей произвольного диктора. Это достигается за счет обобщения отдельных элементов, включая артикуляционные особенности произнесения звуков, в единое, что удачно сочетается с технологией HTM, используемой в Numenta.

Таким образом, показано распознавание гласных звуков от неизвестного диктора (Рис. 3-8) и результаты распознавания их в системе Numenta (Рис. 9-14).

На основании сказанного можно сделать вывод, что система (Рис.1) позволит в ближайшем будущем комплексно решить проблему распознавания слитной речи от неизвестного диктора и изображений этих дикторов в рамках единой системы контроля с помощью веб-камер.

В перспективе, как описано в работе [14], будет обеспечен автоматический контроль с распознаванием видеоинформации, слитной речи от неизвестного диктора с пониманием смысла произносимых фраз и действий контролируемых субъектов и объектов.

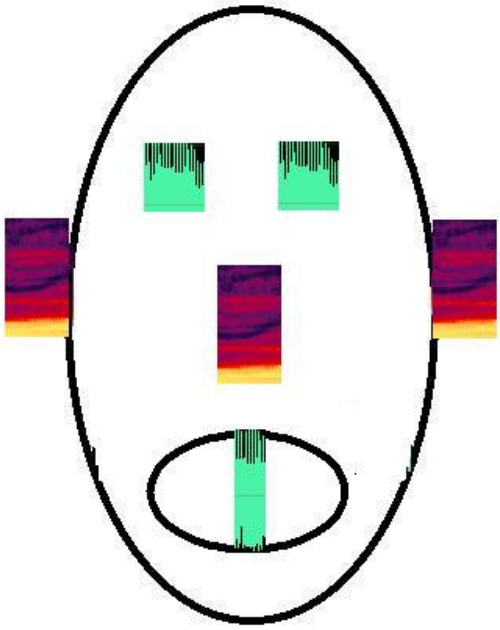

Рис. 3. Физиономический образ звука А.

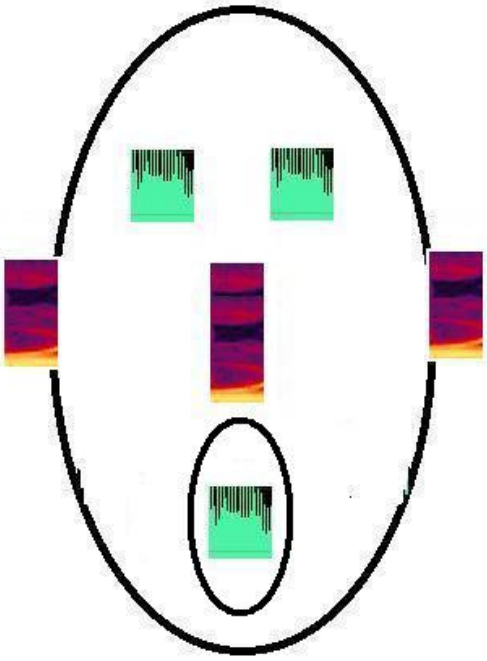

Рис. 4. Физиономический образ звука О.

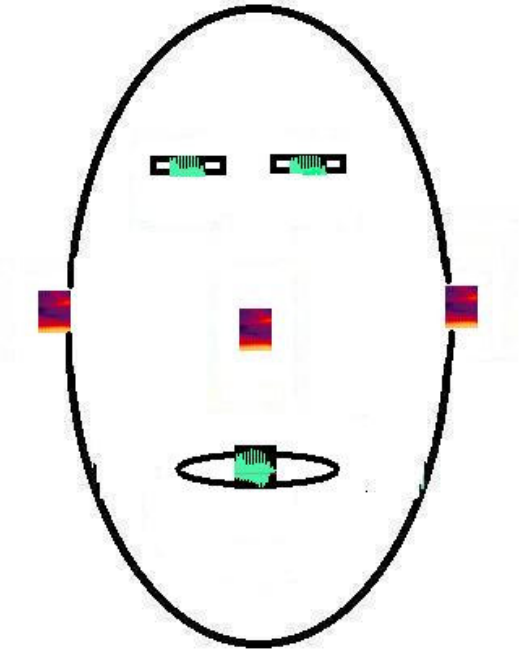

Рис. 5. Физиономический образ звука И.

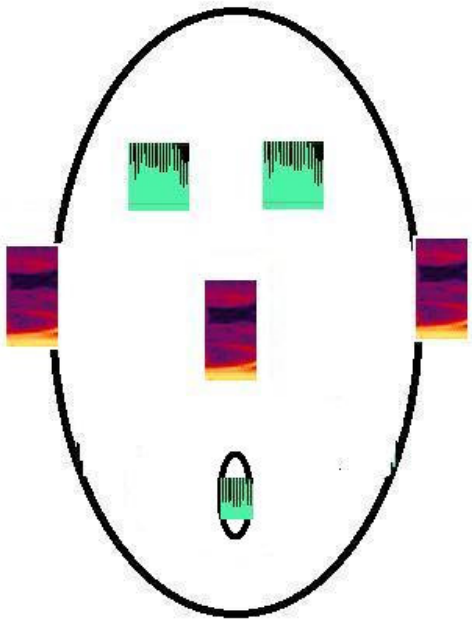

Рис. 6. Физиономический образ звука У.

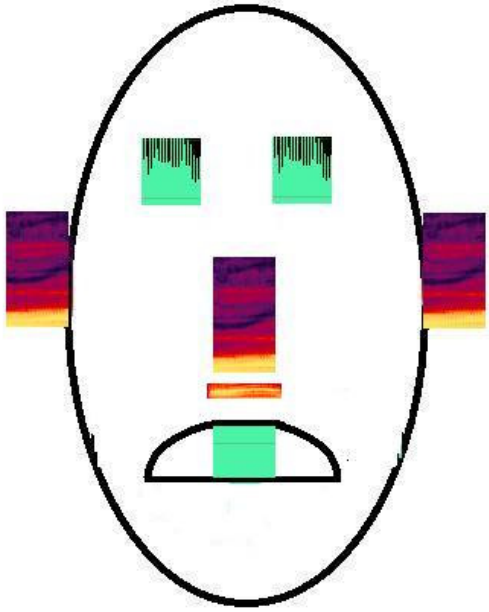

Рис. 7. Физиономический образ звука Э.

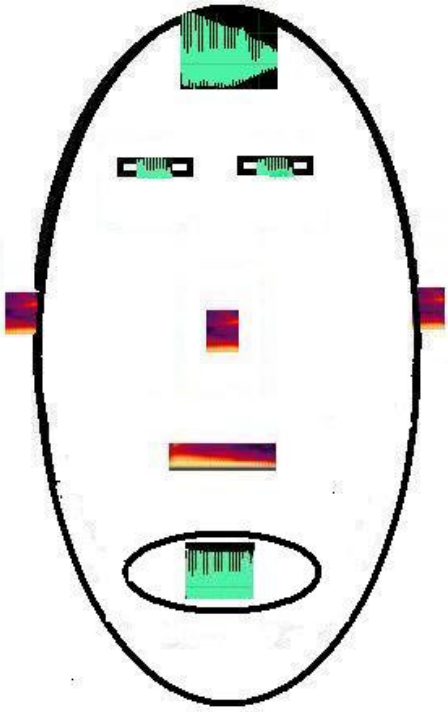

Рис. 8. Физиономический образ звука Ы.

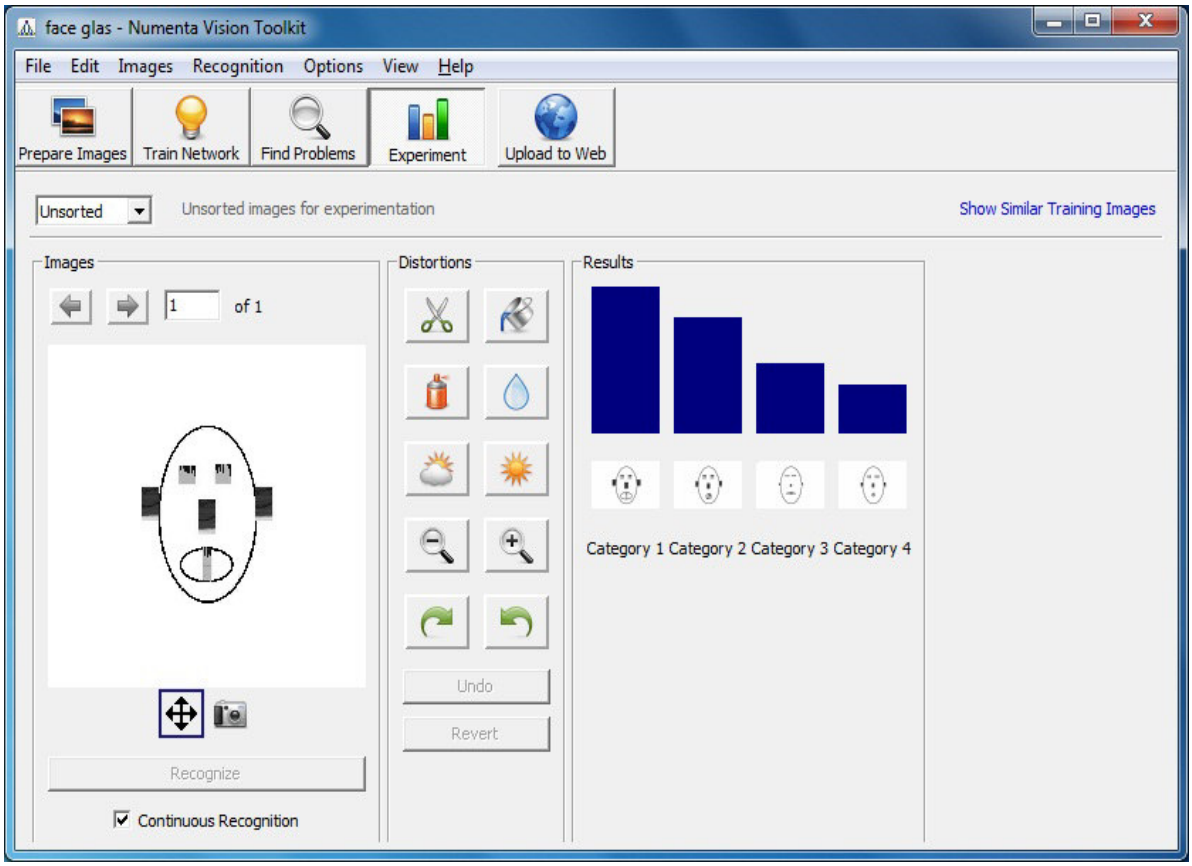

Рис. 9. Распознавание звука А.

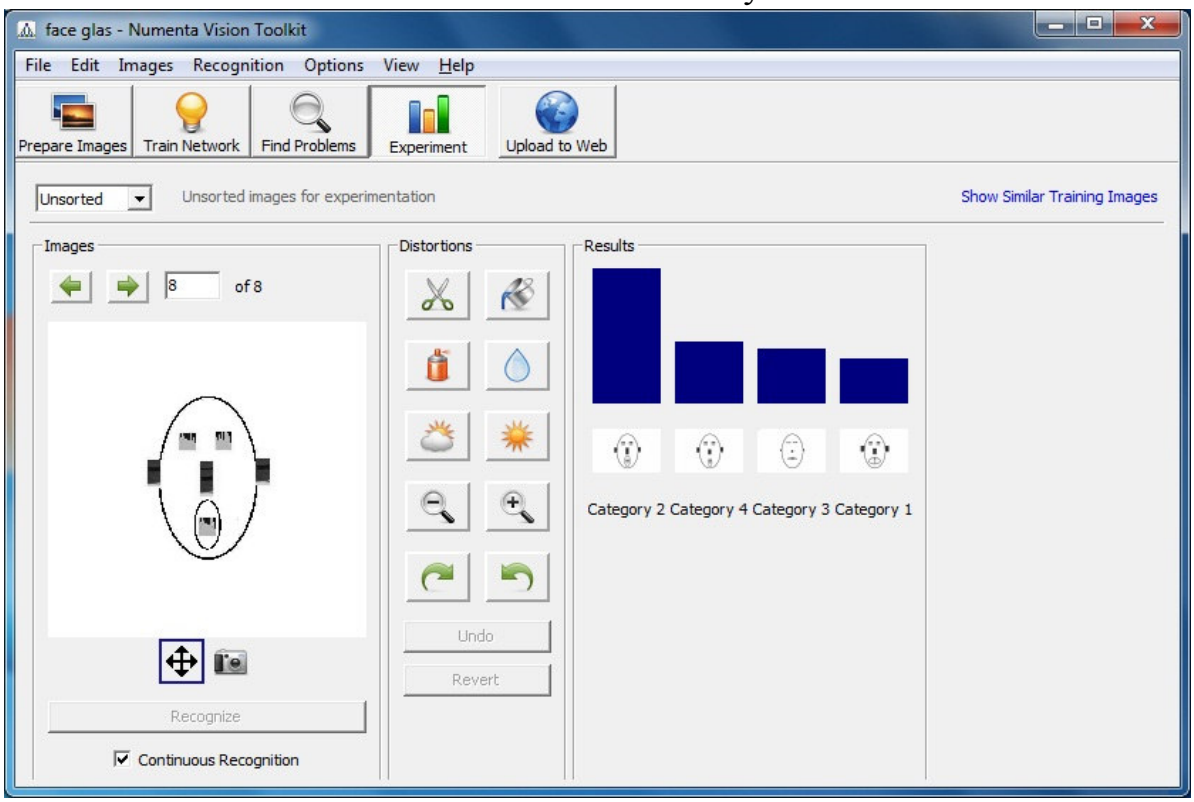

Рис. 10. Распознавание звука О.

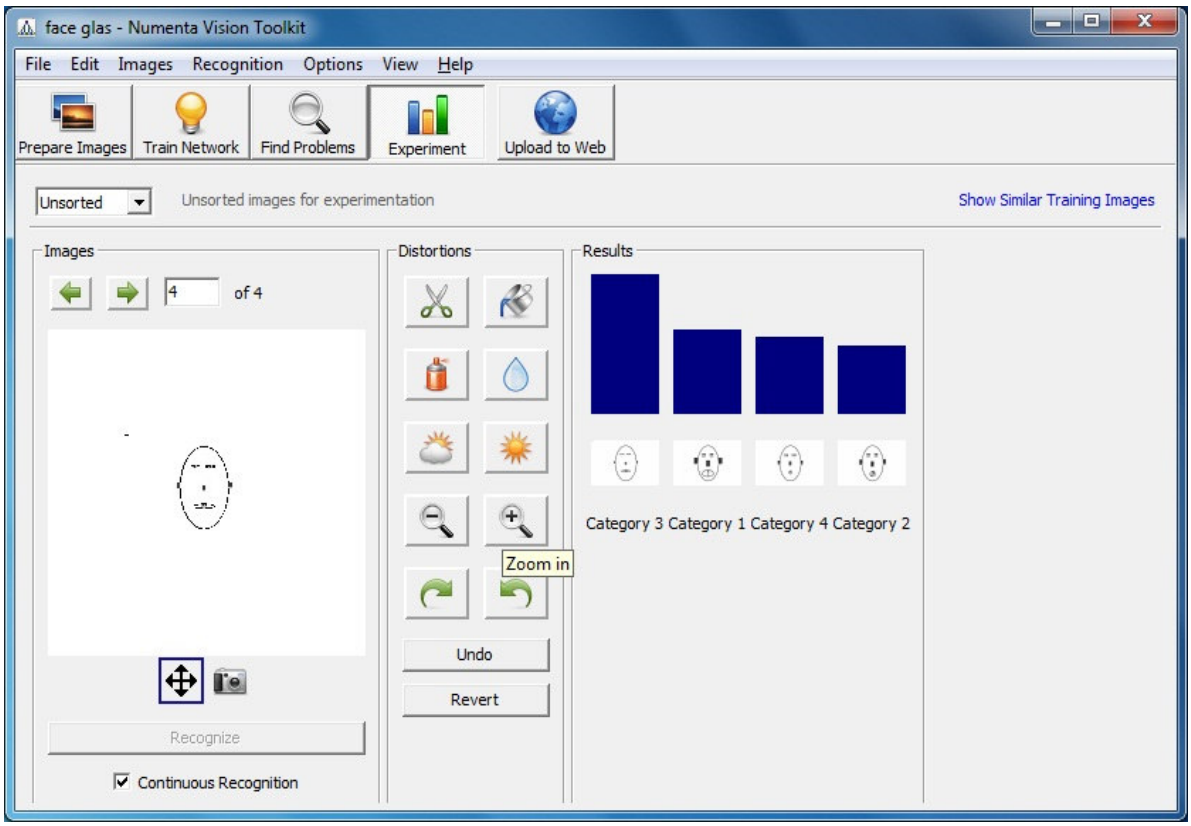

Рис. 11. Распознавание звука И.

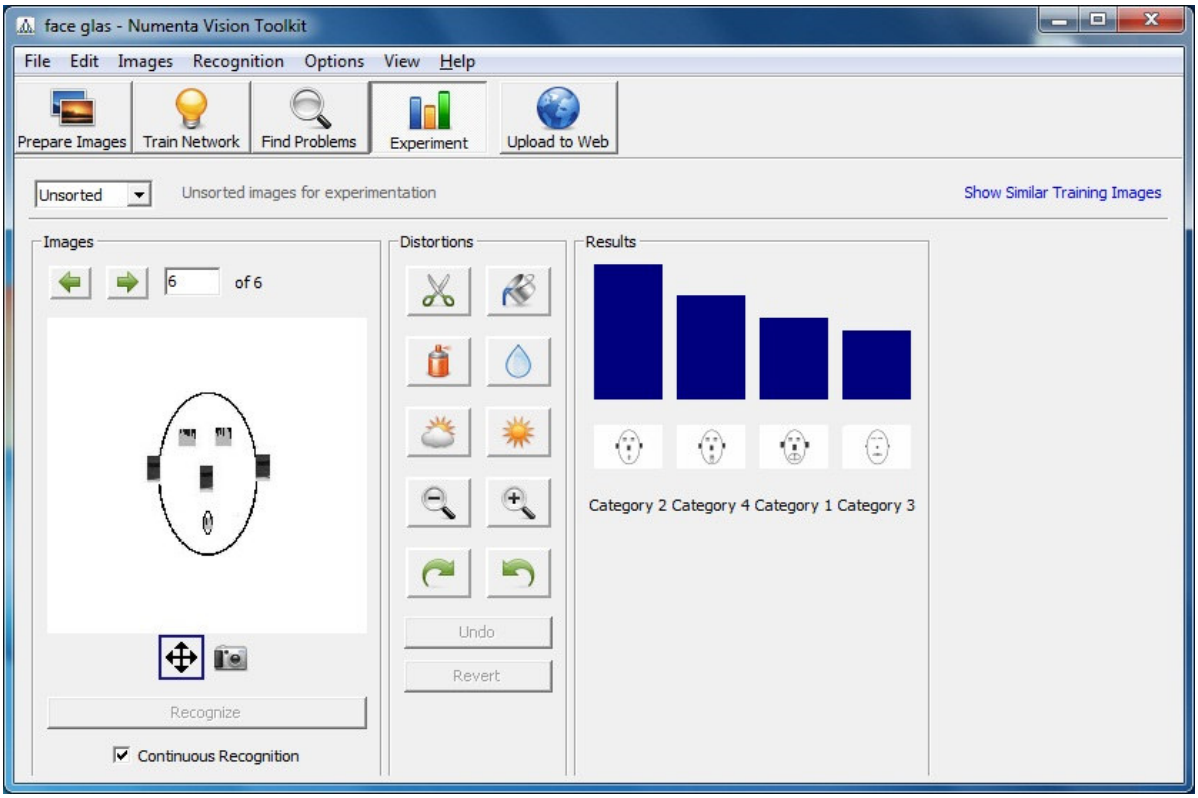

Рис. 12. Распознавание звука У.

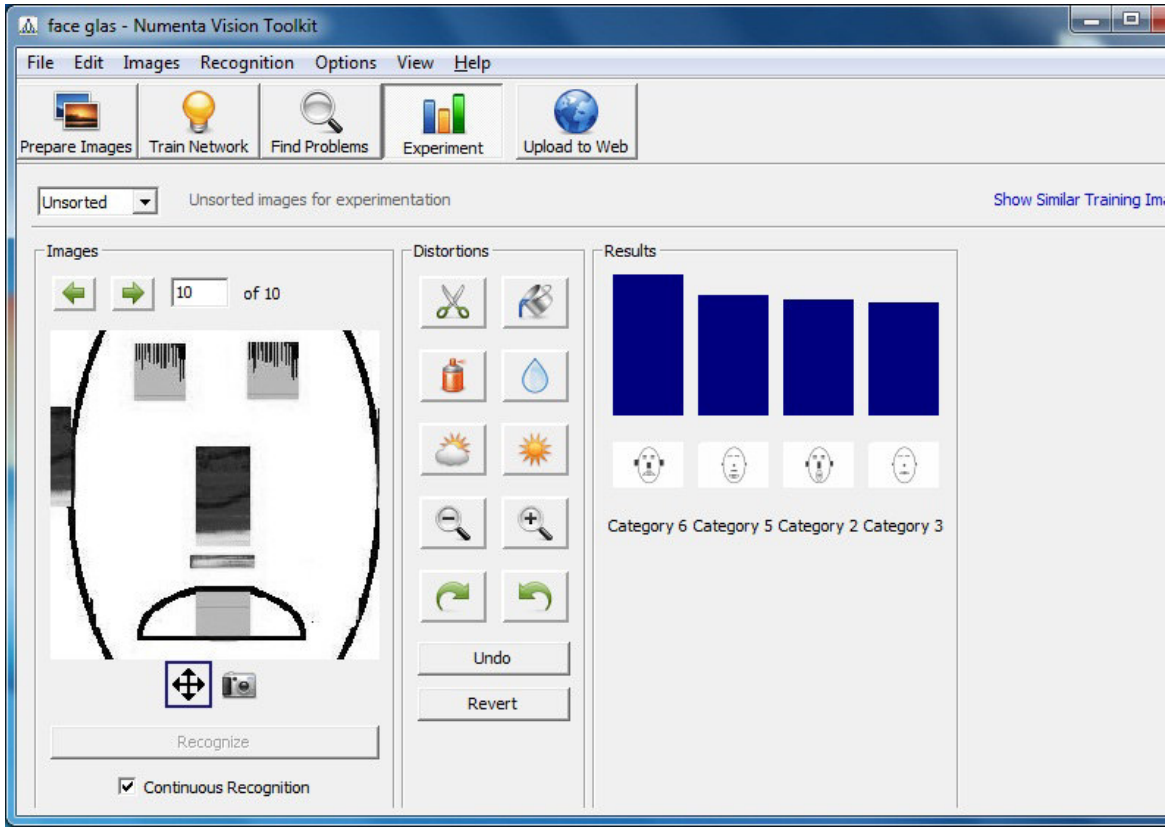

Рис. 13. Распознавание звука Э.

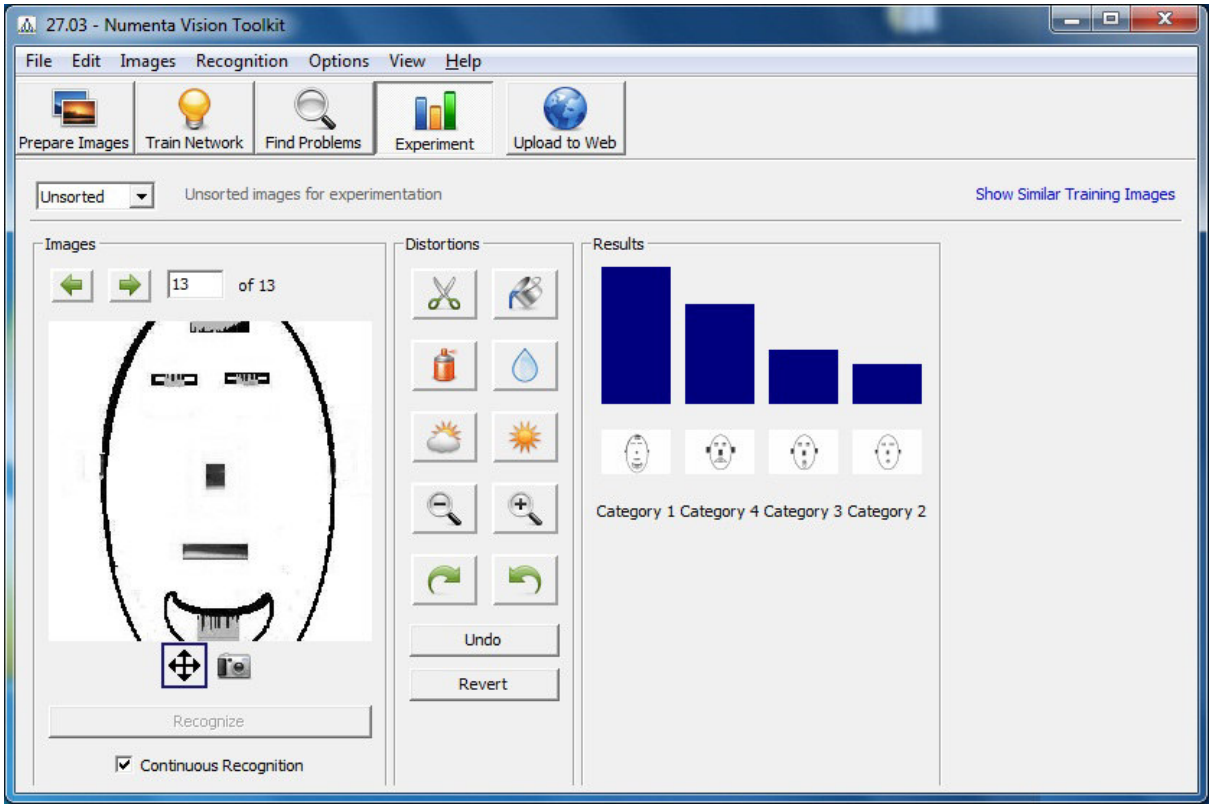

Рис. 14. Распознавание звука Ы.

Литература

1. Потапов А.С. Распознавание образов и машинное восприятие. Спб.: Политехника, 2007. - 548с.

2. Насыпный В.В. Развити е теории построения открытых систем на основе информационной технологии искусственного интеллекта. М.: Воениздат, 1994. - 248с.
3. Насыпный В.В., Насыпная Г.А. Способ синтеза самообучающейся системы извлечения знаний из текстовых документов для поисковых систем. Патент РФ №2273879, номер международной заявки PCT/RU02/00258, дата подачи 28 мая 2002.
4. Насыпный В.В., Насыпная Г.А. Способ синтеза самообучающейся аналитической вопросно-ответной системы с извлечением знаний из текстов, заявка на патент №2007120344/09 от 06.08.2007. Получено решение на выдачу патента на изобретение от 21.07.2008.
5. Современный русский язык: Учеб.дляфилол. спец. высших учебных заведений. Под редакцией В.А. Белошапковой. М.: Азбуковник, 1999. – 928с.
6. Насыпный В.В., Насыпная Г.А. Система распознавания, понимания смысла, анимационного моделирования и синтеза речи на основе стохастической информационной технологии. М.: Прометей, 2008. – 76 с.
7. Искусственный интеллект. Справочник. Кн. 2. Модели и методы. Под ред. Поспелова Д.А. М.: Радио и связь, 1990. - 303 с
8.Насыпный В.В. Сохастика как основа для перехода к большим данным, индустрии знаний и нанотехнологии. – М.: МПГУ. 2011. – 24 с.
9. Насыпный В.В. Способ комплексной защиты распределенной обработки информации в компьютерных системах и система для осуществления способа. Патент РФ №2259639, номер международной заявки PCT/RU /00272, дата подачи 28.10.2003г.
10. Насыпный В.В., Насыпная Г.А. Метод семантической связи текста с трехмерной графикой. – М.: Прометей, 2007. – 27с.
11. Галунов В.И., Чистович Л.А. О связи моторной теории с общей проблемой распознавания речи. Акустический ж., т. 11, с.417-426.
12. Сорокин В.И. Моторная теория восприятия речи и теория внутренней модели // В сб.: Информационные процессы, ИПИ РАН. Том 7. 2007, №1, с.1-12.
13. Halsall F. Data communications computer networks and osi. Addison-wesley publishing company, 1988. - 973 с.
14. Насыпный В.В. Распознавание и понимание смысла речи на основе стохастики в шумах. М.: Прометей, 2010. – 139 с.

7. Аналитика и поиск

Концептуальные основы построения самообучающихся аналитических

систем с извлечением знаний из текстов по различным тематическим областям

Для реализации описанного в предыдущем разделе процесса распознавания, понимания и синтеза речи предложен программный комплекс интеллектуальных систем. Этот комплекс включает самообучающуюся аналитическую систему с извлечением знаний из текстов, а также интеллектуальные системы анализа и синтеза речи. В данном разделе рассмотрим концептуальные основы построения самообучающейся аналитической системы, которая предназначена, прежде всего, для семантического анализа с целью определения смысла распознаваемой слитной речи от неизвестных дикторов.

Отметим, что без понимания смысла речи, как было показано выше, невозможно достоверное распознавание речевого сигнала. При этом система должна обеспечивать семантический анализ отдельных лексических единиц, а также словосочетаний, предикатов, предложений и абзацев текста. Это обеспечивает смысловое сопровождение процесса распознавания речи и выделения ключевых слов в рамках определенной тематики описываемых событий или действий, выраженных в смысловом контексте формируемого текстового представления речи. Для этого требуется эффективная интеллектуальная обработка с использованием больших объемов знаний и реализации логического вывода в реальном масштабе времени при шумовом воздействии. Отметим, что современные интеллектуальные системы не обеспечивают решения указанных задач ввиду эффекта «комбинаторного взрыва». Как показано в работе [2], эта задача успешно решается на основе стохастической информационной технологии.

В данном разделе описан порядок построения и применения самообучающихся интеллектуальных аналитических систем с извлечением знаний из текстов для понимания смысла речи. Эти изделия подробно описаны в [3, 4].

Как было отмечено выше, указанные системы создаются на основе стохастической информационной технологии, разработанной в России. Цель - построение на базе современного компьютера (машины Тьюринга) нового виртуального компьютера для эффективной лингвистической, семантической и логической обработки текстов.

Выбор тематики аналитических систем определяется содержанием неструктурированной текстовой информации, полученной в ходе смыслового анализа распознаваемого речевого сигнала. При этом аналитические функции, реализуемые в системе, которые связаны с индуктивным и дедуктивным логическим выводом, аналогией, обобщением, сравнением и др., широко применяются в ходе семантического анализа распознаваемого текста. Отметим, что при самообучении системы происходит формирование «картины мира» и системы семантической классификации понятий, словосочетаний и предикатов, входящих в состав «картины мира», без которых не возможен полноценный семантический анализ текстов.

Отметим, что данная система обеспечивает возможность извлечения знаний из речевых образов, при этом описание параметров и характеристик речевых сигналов, как было показано выше, автоматически переводится в текстовый вид и представляется в виде соответствующих предикатов и словосочетаний. Указанные предикаты и словосочетания содержат необходимые классификационные параметры различных звуков, а также описание дополнительных характеристик.

Первым уровнем обработки после выделения лексемы из слитной речи является его морфологический анализ. На втором уровне проводится синтаксический анализ, который реализуется с помощью специальной базы знаний, представленной в виде правил продукций, обеспечивающие синтаксический разбор простых и сложных предложений текста. При этом в лингвистический индекс каждого слова заносятся соответствующие синтаксические коды, определяющие данное слово как член предложения.

Семантический анализ текста проводится параллельно с синтаксическим и начинается с автоматически выполняемой классификации общего словаря и специальных толковых словарей терминов и определений по заданным предметным областям, которые связаны с тематикой данной аналитической системы.

При классификации активно используются аналитические функции индуктивного и дедуктивного анализа и синтеза связи слов, обрабатываемых в толковых словарях. В результате образуются семантические классификаторы, представленные в виде таблиц. Входом в таблицы являются стохастические индексы основ слов, строки таблицы содержат иерархию подклассов каждого слова и конечный класс, к которому данное слово принадлежит. Поскольку классификатор сделан для всех частей речи словарей, он позволяет определять типы, а также подклассы и классы объектов и связей между ними.

С помощью классификатора формируются правила продукций семантического анализа текста, которые записываются в специальную базу знаний. После проведения пословного семантического анализа лингвистический индекс каждого слова дополняется его семантическими характеристиками. В результате этого завершается процедура лингвистического анализа текста, после которого каждое слово каждого предложения будет представлено двумя стохастическими индексами: уникальным стохастическим индексом – идентификатором и лингвистическим индексом данного слова, содержащего все его морфологические, синтаксические и семантические характеристики, необходимые для дальнейшей индексации и разбора.

После этого переходят к построению таблицы индексов данного текста в составе локальных, корпоративных баз данных или сайтов Интернет. Левый столбец таблицы содержит индексы неповторяющихся основ слов, входящих в текстовые документы по данной тематике, а строки содержат лингвистический индекс и адресную часть в виде совокупности индексов названия текстовых документов, индекса абзаца, предложения и предиката, в

котором содержится данный индекс слова. Таблицы индексов текста используются при первичном поиске ответов или необходимых предложений текста с применением ключевых слов. Поиск по ключевым словам является основой для реализации второго уровня поиска с использованием семантики, извлечения знаний из текстов и аналитики.

Затем переходят к формированию концептуального описания предметной области текстов на основе выделенных в стохастической форме предикатов. Концептуальное описание представляется также в виде таблицы. Левый столбец содержит стохастические индексы всех неповторяющихся словосочетаний и предикатов индексируемого текста, строки включают индексы типов объектов и отношений между ними, а также (с использованием классификаторов) соответствующие им классы. Кроме этого, в состав таблицы также входит адресная часть, включающая индексы текста, абзаца и предложения, куда входят предикаты, которые содержат указанные классы объектов и отношений между ними. Это позволяет, используя классификатор и концептуальное описание предметной области, производить более точный повторный поиск необходимой информации после выполнения поиска по ключевым словам с тем, чтобы более полно и точно находить необходимые ответы или предложения, используя близкие по смыслу слова, словосочетания и предикаты, активно применяя семантический анализ текста.

На основе сформированного концептуального описания предметной области текста, а также используя формализованное описание функций определения, обобщения, сравнения, выбора, аналогии, дедукции и индукции, анализа и синтеза автоматически формируются правила продукций, содержащие необходимые типы и классы логически связанных предикатов предметной области текста. На основе этих функций могут формироваться деревья логического вывода, содержащие необходимые комбинации исходных логических функций, которые требуются пользователю системы для получения результата аналитического анализа с целью формирования обобщенных семантических характеристик словосочетаний, предикатов и сформированных из них предложений текста. Отметим, что текст, формируемый после распознавания вводимых речевых сигналов, может также использоваться для эволюционного развития описания предметной области – «картины мира». При этом для повышения эффективности распознавания речи введение аналитического поиска существенно увеличивает полноту поиска и обработки информации исходных текстовых файлов или сайтов. Это обусловлено тем, что непосредственно к декларативной составляющей текстовых баз добавляются новые знания, извлекаемые из текста с помощью базовых аналитических функций и их заданных комбинаций. За счет комбинаций базовых функций исходная аналитическая система может автоматически настраиваться на заданную предметную область и эффективно использоваться в той области, к которой

относится вводимая речевая информация: например, управление, социальное обеспечение, финансирование, образование, культура, спорт и другие.

Для извлечения знаний из больших объемов неструктурированных текстов различных типов (диссертации, монографии, учебно-методическая, справочно-энциклопедическая литература и др.), которые связаны с проблемой распознавания речи с использованием описанного выше многоуровневого анализа речевых и текстовых сообщений аналитическая система может работать в автоматическом вопросно-ответном режиме. Здесь могут применяться разные варианты работы, например, осуществление точного семантического поиска, если информация непосредственно содержится в тексте и может быть выдана по запросу.

В более сложных случаях автоматически реализуются аналитические функции, которые после предварительной обработки информации с использованием процедур логического вывода, эквивалентных преобразований дают ответы на поставленные вопросы. Доказано, что если в системе может быть синтезирован алгоритм, который выдает ответ на поставленный вопрос с применением индексированной текстовой базы, то может быть создан аналитический алгоритм с использованием комбинаций разных функций, который обеспечит представление пользователю заданной информации.

В результате повышается эффективность формирования «картины мира» и обеспечивается полнота представленных понятий и связей между ними. На основе полученных предикатов, входящих в «картину мира», автоматически формируются правила продукций по различным проблемным областям, в том числе и по проблеме распознавания речи. В этом случае между предикатами семантической сети, которые отображают «картину мира», выделяются семантические связи типа «условие-заключение», «причины-следствия», цели, определения и другие.

Как известно, правила продукций представляют собой символьную конструкцию вида «если (условие), то (заключение)». При этом условия содержат совокупность предикатов, объединенных логическими связками «и», а заключение содержит предикат, который выполняется, если все предикаты, входящие в условие, являются истинными для какой-то конкретной ситуации, соответствующей исследуемым объектам или процессам в определенной области знаний, например, при распознавании речи. Все полученные правила автоматически проверяются на их смысловую корректность. После стохастической индексации записываются в базы знаний. Таким образом, производится наполнение всех отмеченных баз знаний, используемых в комплексном процессе понимания текста.

Представление словосочетаний, предикатов «картины мира» и правил продукций в стохастически индексированном виде дает возможность использовать эффективные алгоритмы логического вывода, а также (с помощью стохастической информационной технологии) исключить проблему «комбинаторного взрыва». Без решения этой проблемы построение описанной выше системы распознавания с использованием многоуровневого

анализа, понимания смысла и синтеза речи в принципе невозможно. Кроме этого стохастическая информационная технология, позволяет эффективно реализовывать функции помехозащищенности и нормализации речи [6].

Литература

1. Потапов А.С. Распознавание образов и машинное восприятие. Спб.: Политехника, 2007. - 548с.
2. Насыпный В.В. Развитие теории построения открытых систем на основе информационной технологии искусственного интеллекта. М.: Воениздат, 1994. - 248с
3. Насыпный В.В., Насыпная Г.А. Способ синтеза самообучающейся системы извлечения знаний из текстовых документов для поисковых систем. Патент РФ №2273879, номер международной заявки PCT/RU02/00258, дата подачи 28 мая 2002.
4. Насыпный В.В., Насыпная Г.А. Способ синтеза самообучающейся аналитической вопросно-ответной системы с извлечением знаний из текстов, заявка на патент №2007120344/09 от 06.08.2007. Получено решение на выдачу патента на изобретение от 21.07.2008.
5. Современный русский язык: Учеб.для филол. спец. высших учебных заведений. Под редакцией В.А. Белошапковой. М.: Азбуковник, 1999. – 928с.
6. Насыпный В.В., Насыпная Г.А. Система распознавания, понимания смысла, анимационного моделирования и синтеза речи на основе стохастической информационной технологии. М.: Прометей, 2008. – 76 с.
7. Искусственный интеллект. Справочник. Кн. 2. Модели и методы. Под ред. Поспелова Д.А. М.: Радио и связь, 1990. - 303 с.
8. Halsall F. Data communications computer networks and osi.Addison-wesley publishing company, 1988. - 973 с.
9. Насыпный В.В. Способ комплексной защиты распределенной обработки информации в компьютерных системах и система для осуществления способа. Патент РФ №2259639, номер международной заявки PCT/RU /00272, дата подачи 28.10.2003г.
10. Насыпный В.В., Насыпная Г.А. Метод семантической связи текста с трехмерной графикой. – М.: Прометей, 2007. – 27с.
11. Галунов В.И., Чистович Л.А. О связи моторной теории с общей проблемой распознавания речи. Акустический ж., т. 11, с.417-426.
12. Сорокин В.И. Моторная теория восприятия речи и теория внутренней модели // В сб.: Информационные процессы, ИПИ РАН. Том 7. 2007, №1, с.1-12
13. Марков А.А. Об одном применении статистического метода // Известия АН, 1916, сер.6, X, №4, с.239-
14. Elinek F. Распознавание непрерывной речи статистическими методами // ТИИЭР 64, 1976, №4, с.131-160.
15. Elinek F. Разработка экспериментального устройства, распознающего раздельно произнесенные слова // ТИИЭР 73, 1985, №11, с.91-99.

16. Галунов В.И. Помехоустойчивость как системообразующий фактор речи // Проблемы и методы экспериментально-фонетических исследований, 2002, с.205-300.
17. Галунов В.И. Речь как система // Труды XIII сессии РАО, 2003, т.3, с.19-21.
18. Kraft D. Speechperception // J. Phonetics, 1979, 7, p.279-312.
19. Галунов В.И., Соловьев А.Н. Современные проблемы в области распознавания речи
20. Liedtke C.-E., Buckner J., Grau O. et al. AIDA: A system for the knowledge based interpretation of remote sensing data // 3d Airborne Remote Sensing Conference and Exhibition. – 1997. – Vol.2. – P. 313-320.
21. Бабин Д.Н., Холоденко А.Б. Использование лексических анализаторов в распозновании образов // Труды международного семинара диалог – 99, Таруса, 1999.

8. Защита

Защищенные стохастические системы

Возможно ли уже сегодня создать компьютерную систему с комплексной защитой от программных закладок, вирусов, действий хакеров и обеспечить информационную безопасность локальных, региональных систем и всей Internet в целом? Как превратить Сеть с ее огромным объемом хаотично циркулирующей информации в глобальную инфраструктуру обработки знаний и их доставки по запросам, сформулированным на естественном языке в виде текста или речи? Допустимо ли решение перечисленных задач в комплексе и при этом на базе единой технологии? Применение оригинальной стохастической технологии для защиты современных компьютеров от всех видов киберинфекции позволяет надеяться, что можно дать положительные ответы на эти вопросы.

Предлагаемая методика основана на введении стохастичности в вычислительный процесс с использованием одноразовых систем шифрования программ и данных при их обработке, хранении и передаче. В основу положена идея адаптации процесса обработки символьной информации к вычислительной среде компьютера [1]. Данный процесс обеспечивается путем стохастического преобразования и кодирования символьных выражений и конструкций. В результате указанные элементы преобразуются в уникальные стохастические индексы — двоичные комбинации заданной длины. Преобразование символьных выражений и конструкций позволяет оптимизировать обработку знаний, данных и текстовой информации путем использования для их представления стохастически индексированных форматов фреймов и реляционных таблиц.

Быстрое построение траектории логического вывода производится за счет непосредственного использования стохастических индексов и кодов с

целью произвольного доступа и обработки семантически связанной информации. В новых интеллектуальных системах, построенных на основе стохастических методов, осуществляется реализация логического вывода на значительных объемах данных, знаний и текстовых документов [1]. При этом обеспечивается линейная зависимость времени логического вывода от объема обрабатываемых данных и знаний. Тем самым решается давняя проблема «комбинаторного взрыва» при логическом выводе на больших объемах информации, которая до сих пор сдерживала развитие интеллектуальных систем, в том числе и поисковых, обеспечивающих извлечение знаний из текстовых документов.

Одновременно указанные преобразования реализуют функцию шифрования исходной символьной информации. Применяются стохастические табличные системы скоростного кодирования, формирования одноразовых секретных и открытых ключей [2]. В результате обеспечивается стохастичность всего вычислительного процесса за счет случайного преобразования полученных таблиц, индексов и кодов после каждого цикла обработки. Семантическое же значение зашифрованных программ, данных и знаний сохраняется на основе принципиально нового метода формирования открытых ключей с использованием аппарата искусственного интеллекта. Для восстановления исходного значения зашифрованной информации в любой заданный момент времени ключ расшифрования вычисляется путем логического вывода на множестве применяемых ключей. Это позволяет избежать необходимости хранения всего множества используемых ключей, а держать в системе лишь начальный ключ и текущий результат логического вывода. Таким образом, обеспечивается гарантированная возможность расшифровки информации, зашифрованной в любое предшествующее время, начиная от первого пуска системы [3].

Выполнение программ, обработка знаний и данных может производиться в зашифрованном виде на основе единого метода логического вывода. Возможности быстрого логического вывода используются и для функций контроля корректности программ путем их верификации. Разработанные методы верификации, основанные на аппарате искусственного интеллекта и стохастического преобразования, обеспечивают гарантированное обнаружение любых программных закладок и вирусов [4]. При этом обеспечивается защита от вновь созданной киберинфекции, включая саморасшифровывающиеся полиморфные вирусы.

Типовая обработка программ и данных заменяется обработкой случайных зашифрованных индексов и кодов, значение которых постоянно обновляется за счет использования одноразовой системы шифрования на основе нового метода формирования открытых и секретных ключей [2, 3]. После выполнения указанного преобразования компьютерные системы приобретают совершенно новые качества. Одно из них заключается в том, что программы и данные обрабатываются, хранятся и передаются только в зашифрованном виде. Как следствие этого, внедряемые программные

закладки и вирусы не могут найти точку входа в программу и воздействовать на нее.

Для повышения стойкости выполняемых программ система комплексной защиты информации реализует два уровня — логический, на основе стохастического преобразования управляющей структуры программы, а также физический, реализуемый за счет стохастического кодирования машинных команд. В ходе обработки обеспечивается контроль корректности каждой логической траектории выполнения программы. При этом осуществляется гарантированная, с заданной вероятностью ошибки способность обнаружения попыток искажения информации и коррекции программных средств и данных [4].

В рамках технологии стохастических кодов впервые решена важнейшая для современных компьютерных систем задача обеспечения обработки числовой информации в защищенном виде при выполнении арифметических вычислений. Решение этой проблемы достигается путем стохастического преобразования и стохастического индексирования информации и сведения арифметических действий к логическим операциям с символьной информацией с использованием стохастически изменяемых таблиц [3]. При применении стохастической информационной технологии обеспечивается высокая степень защищенности программ и данных в случае попытки несанкционированного доступа к их содержанию. Она определяется необходимым числом переборов $N > 10^{100}$ на множестве случайных кодов команд и данных. Так как в процессе функционирования коды программ и данных динамически изменяются, то оценки числа переборов верны для случая анализа «мгновенного среза процесса», т.е. состояния приостановки выполнения стохастически преобразованной программы.

В ходе выполнения программ с использованием данных в зашифрованном виде обеспечивается определенный порядок их взаимодействия. Операционная система, функционирующая в защищенном виде, управляет исполнением зашифрованной прикладной программы, производящей обработку криптографически защищенных данных. Последующая реализация новой стохастической информационной технологии связана с ее внедрением в компьютерные системы на более глубинных уровнях. Речь идет об использовании новых технологий при создании операционных систем, а также об аппаратном воплощении (создание специальных процессоров). Здесь образуется единый технически замкнутый контур шифрования информации при ее обработке, хранении и передаче. Это обеспечит существенное повышение защищенности программ и данных на физическом уровне (коды команд и данных). В результате формируется защищенная программно-аппаратная среда, включающая модернизированные аппаратные средства, обеспечивающие необходимое преобразование и шифрование информации, а также криптографически защищенные операционную систему, прикладные программы, базы данных и знаний. При этом выполнение программ и обработка информации в

зашифрованном виде сопровождаются функциями контроля и поддержания их корректности и целостности.

На основе указанных идей разработана новая технология [2, 3], позволяющая: создавать защищенные процессоры; разрабатывать высокоскоростные кодеры/декодеры, производительность которых сопоставима с производительностью типового процессора компьютера; создавать систему формирования и передачи открытых и закрытых ключей, блоков получения электронно-цифровой подписи и хэш-функции; строить защищенные арифметические процессоры для обработки в зашифрованном виде чисел повышенной разрядности; создавать устройства стохастического перекодирования («перекодеры»), позволяющие перешифровывать программы и данные, используя различные ключи без раскодирования информации; формировать защищенные программно-аппаратные интерфейсы на базе перекодеров, обеспечивающих взаимодействие программ и данных в защищенном виде.

Все это позволяет реализовать в компьютерной системе полностью закрытый контур обработки программ и данных в защищенном виде. Данный контур реализуется на уровне кодов машинных команд с использованием как типовых, так и вновь созданных устройств обработки и защиты информации. Создание контура защиты прозрачно для системы в целом и не изменяет ни ее концепции, ни функций администрирования, ни управления вычислительным процессом, ни протоколов взаимодействия программ и данных. Описанная стохастическая технология на базе одноразовых систем шифрования может быть реализована не только в рамках отдельных компьютеров, а и в вычислительных средствах и каналах открытых систем [2]. За счет этого возможно достижение гарантированной защищенности в любых современных компьютерных сетях, включая Internet.

Литература

1. Насыпный В.В. Развитие теории построения открытых систем на основе информационной технологии искусственного интеллекта. М.: Воениздат, 1994.
2. Насыпный В.В. Одноразовое шифрование с открытым распределением ключей. // Открытые системы. 2004, № 1.
3. Насыпный В.В. Способ комплексной защиты распределенной обработки информации в компьютерных системах и система для осуществления способа. Международная заявка на изобретение № PCT/RU 01/00272 от 05.07.2001.
4. Насыпный В.В. Комплексная защита компьютерных систем. // Мир ПК, 1998, № 4.

9. Шифрование

Одноразовое шифрование с открытым распределением ключей

Одноразовую систему шифрования разработали еще в 1917 году Дж. Моборн и Г. Вернам [2]. Ее характерная особенность — одноразовое исполь-

зование ключевой последовательности. Такая система шифрует исходный открытый текст X в шифротекст Y с использованием одноразовой случайной ключевой последовательности K. Для ее реализации иногда используют одноразовый блокнот, составленный из отрывных страниц; на каждой из них напечатана таблица со случайными числами (ключами) Ki. Блокнот выполняется в двух экземплярах: один используется отправителем, а другой — получателем. Для каждого символа Xi сообщения имеется свой ключ Ki из таблицы получателя. После того, как таблица использована, ее необходимо удалить из блокнота и уничтожить. Шифрование нового сообщения начинается с новой страницы.

Обсолютная надежность одноразовой системы доказана Клодом Шенноном в его известной работе «Теория связи в секретных системах», отрывок из которой приведен в [3]. Одноразовые системы нераскрываемы, поскольку их шифротекст не содержит достаточной информации для восстановления открытого текста. Однако возможности использования одноразовых систем на практике ограничены. Ключевая последовательность длиной не менее длины сообщения должна передаваться получателю сообщения заранее или отдельно по некоторому секретному каналу, что практически неосуществимо в современных информационных системах, где требуется шифровать многие миллионы символов и обеспечивать засекреченную связь для множества абонентов. Эти недостатки устранены в способе синтеза одноразовых систем шифрования с открытым распространением ключа [5].

Рассмотрим процесс передачи информации по линии связи, соединяющей пользователей A и B. Предлагаемый способ построения одноразовой системы дает возможность передавать практически неограниченный объем информации с использованием случайной перестановки только одной таблицы ключей.

В качестве базового шифрующего элемента для системы с открытой передачей ключей разработан одноразовый многоалфавитный кодер (ОМК). Система содержит ОМК, датчик случайных чисел, схему формирования случайной перестановки на стороне A и многоалфавитный декодер на стороне B. ОМК реализует процесс стохастического кодирования [4].

В состав ОМК входит базовая таблица одноразовых ключей, регистр перестановки интерфейса, регистры случайной и псевдослучайной перестановок строк и столбцов базовой таблицы. Аналогичный состав имеет и многоалфавитный декодер. Регистры случайных и псевдослучайных перестановок строк и таблицы интерфейса декодера содержат комбинации, обратные по отношению к соответствующим перестановкам кодера.

Базовая таблица одноразовых ключей на стороне A и на стороне B имеет размер n x n. Каждая i-я строка таблицы содержит случайную ключевую комбинацию, в которую входят все возможные различные значения Kij длиной m бит. (Для таблицы кодов ASCII m = 8, n = 256, поэтому для шифрования текста используют таблицу размером 256 x 256.)

Ki = K i0, Ki1, ..., K in-1 (i = 1,... n)

В результате работы датчика случайных чисел и схемы формирования случайной перестановки генерируется соответствующая перестановка. В полученной перестановке столбцы задают соответствие между входными значениями (верхняя строка) и выходными (нижняя строка).

Базовая таблица одноразовых ключей на стороне А выполняет две функции:

генерацию виртуальной переменной таблицы одноразовых ключей со случайной перестановкой столбцов и строк;

реализацию логического вывода, обеспечивающего преобразования секретной перестановки в несекретную, применяемую для открытой передачи ключа.

С этой целью каждый столбец базовой таблицы можно представить в виде вертикально расположенной перестановки. При этом регистр псевдослучайной перестановки, подключенный к данной таблице, в сочетании с предыдущей случайной перестановкой, которая передана на сторону В, обеспечивает выбор столбцов таблицы для формирования их одноразовых комбинаций. Названные комбинации столбцов применяются в процессе логического вывода. Всего может быть сформировано N = n! различных комбинаций столбцов. Логический вывод реализует однонаправленную функцию Y = F(x), которая позволяет на основе секретной перестановки, записанной в левый регистр базовой таблицы одноразовых ключей, получить несекретную перестановку, формируемую в выходном блоке ОМК. Здесь x — значение секретной перестановки, F — функциональные связи, формируемые в процессе логического вывода с использованием очередной комбинации столбцов-перестановок, Y — относительная несекретная перестановка. Зная x и формируя функциональные связи F, легко получить Y. Однако по известному значению Y, не зная всей схемы функциональных связей базовой таблицы, нельзя восстановить исходную секретную перестановку. Для этого необходимо произвести полный перебор на множестве V = n! всех значений результирующих перестановок, получаемых в ходе логического вывода, — своего рода эффект лабиринта, в центр которого помещают человека с завязанными глазами и, сняв повязку, предлагают путем случайного перебора всех возможных вариантов прохода найти выход.

Таким образом, одновременно с передачей и шифрованием информации на стороне пользователя А генерируется очередная случайная перестановка. Затем с помощью описанного алгоритма логического вывода формируется соответствующая ей несекретная перестановка. Она передается на сторону В в начале обмена информацией и после передачи по линии связи n блоков шифротекста длиной n символов каждый. На основе этой перестановки на стороне В с помощью базовой таблицы, идентичной базовой таблице А, выполняется процедура обратного логического вывода с целью получения соответствующей секретной перестановки . Эта процедура описывается выражением $\mathbf{x = F^{-1}(Y)}$, где F^{1} — функция обратного логического вывода, реализуемого с помощью базовой таблицы стороны В. Сформированная секретная

перестановка записывается в регистры случайных перестановок столбцов и строк многоалфавитного декодера. Путем использования указанных регистров в декодере происходит образование виртуальных таблиц одноразовых ключей в соответствии с полученной случайной перестановкой. В результате на сторонах А и В каждый раз будут одновременно сформированы новые случайные виртуальные таблицы одноразовых ключей, идентичных по содержанию. Эти таблицы применяются при передаче зашифрованной информации.

Рассмотрим этот процесс подробнее. Исходный текст поступает на вход регистра перестановки интерфейса ОМК, который обеспечивает перестановку таблицы кодов ASCII. Так осуществляется первый этап преобразования исходной информации. Затем преобразованный текст проходит через регистр случайной перестановки строк, которая в сочетании со случайной перестановкой столбцов реализует очередную виртуальную таблицу одноразового ключа. При этом применение случайных и псевдослучайных перестановок обеспечивает для каждой очередной комбинации исходного текста Xi = (Xi0, Xi1, ..., Xi,n-1) (i = 1,... n) формирование уникальной одноразовой ключевой последовательности Ki = Ki0, Ki1, ..., Kin-1 (i = 1,... n). Всего для данной виртуальной таблицы, определяемой очередной случайной перестановкой, может быть образовано n таких ключевых последовательностей. В результате произведенных перестановок и замен в многоалфавитном кодере символов каждой очередной последовательности Xi, а также циклических сдвигов столбцов таблицы, процесс шифрования аналогичен классической одноразовой системе. В декодере сначала реализуется процедура идентификации символов шифротекста путем включения соответствующих столбцов базовой таблицы, а затем производятся соответствующие циклические сдвиги столбцов и с помощью регистров перестановок строк выполняются обратные перестановки, обеспечивающие преобразование шифротекста в исходный текст.

После передачи i = n очередных комбинаций шифротекста реализуется описанный процесс открытой передачи ключа (очередной секретной перестановки). За счет этого производится постоянная (с заданным периодом) случайная модификация виртуальной таблицы многоалфавитных кодера и декодера для получения новых таблиц одноразовых ключей. Затем продолжается передача, шифрование и дешифрование информации с использованием новых таблиц одноразовых ключей. При этом передача несекретной перестановки реализует функцию открытой передачи ключей, производимой после выдачи каждых n блоков зашифрованной информации. В результате обеспечивается гарантированная надежность шифрования. Действительно, сами базовые таблицы одноразовых ключей противнику неизвестны при любых видах атак на данную систему шифрования (в явном виде они не участвуют в процессе шифрования информации), поэтому формируемые виртуальные таблицы одноразовых ключей случайны и непредсказуемы. Учитывая однонаправленность функции Y = F(x) получения несекретной перестановки, множество вариантов модификации виртуальных таблиц на сторонах А и В путем случайной перестановки столбцов и строк измеряется числом V = n! Так, при использовании таблицы кодов ASCII с указанными параметрами m и n получим

величину $V > 10^{500}$. Для больших значений n данный способ позволяет передавать практически неограниченные объемы зашифрованной информации в режиме одноразового ключа с гарантированным уровнем надежности, определяемым числом V = n! всех возможных значений результирующих перестановок, которые получают в ходе логического вывода. Отметим, что в данном случае применяется одна таблица одноразовых ключей размером n x n и функция открытой передачи ключей с использованием случайной несекретной перестановки длиной n байт. Трудно даже указать, сколько времени потребуется на переборы всех вариантов перестановок на реальном компьютере. При этом функция открытой передачи ключей может периодически использоваться для обновления базовой таблицы путем передачи новых значений ее столбцов (перестановок). Указанные значения столбцов генерируются с помощью датчика случайных чисел и схемы формирования случайных перестановок. В результате после n циклов обновления на сторонах А и В будут получены новые базовые таблицы, используемые далее при шифровании.

Процесс кодирования в ОМК практически не снижает скорость передачи информации по каналу связи. Это позволяет реализовать скоростные одноразовые шифры для работы в компьютерных сетях. Имеются эффективные технологии обеспечения целостности информации, а также идентификации и аутентификации пользователей, проверки подлинности сообщений.

Система с открытым распределением ключей

В состав системы с открытым распределением ключей [5] входят центр сертификации, формирования и распределения ключей (ЦСФРК), серверы распределенной обработки и пользовательские устройства. В качестве шифрующего элемента применяется ОМК.

Основными задачами ЦСФРК являются подключение пользовательских устройств и серверов к системе защиты, их сертификация, формирование и распределение закрытых и открытых ключей между пользовательскими устройствами и серверами распределенной обработки данных. В ЦСФРК генерируется и хранится главный ключ системы (мастер-ключ), который представляет собой случайно заполненную кодами таблицу размером n x n.

На основе таблицы главного секретного ключа в ЦСФРК путем случайной перестановки столбцов и строк формируется множество различных таблиц начальных секретных ключей для пользователей. При этом каждой полученной таблице начального секретного ключа ставится в соответствие примененная перестановка столбцов и строк таблицы главного секретного ключа.

Затем для каждой таблицы начального секретного ключа путем случайных перестановок его столбцов и строк создаются таблицы внутреннего секретного ключа и внешнего секретного ключа. Каждой полученной таблице ставятся в соответствие использованные случайные перестановки столбцов и строк таблицы начального секретного ключа.

Полученные таблицы начального ключа и случайные перестановки столбцов и строк для формирования таблиц внутреннего секретного ключа, а также внешнего секретного ключа применяются при подготовке носителей для сертифицированных пользователей. Формируется носитель данных — смарт-карта, копия которой хранится в центре сертификации. Она содержит полную таблицу начального ключа, а также набор секретных ключей-перестановок для таблиц внутреннего и внешнего ключей пользователя. Также в смарт-карту записывается PIN-код и значение хэш-функции пароля данного пользователя.

Чтобы получить систему ключей, пользователь вводит информацию со смарт-карты; при доступе к функциям системы защиты по команде пользователя в пользовательском устройстве на основе таблицы начального ключа и секретных перестановок, введенных со смарт-карты, производится формирование таблиц внутреннего секретного ключа, а затем таблицы внешнего секретного ключа. Аналогичные процедуры выполняются и на сервере. При этом таблица внешнего секретного ключа применяется для заполнения базовой таблицы одноразовых ключей ОМК, который служит для организации внешней зашифрованной связи с другими пользователями или серверами сети.

После завершения процесса формирования ключевых таблиц пользователь может обратиться с запросом к ЦСФРК для организации закрытой связи с требуемым сервером распределенной обработки или с другим пользователем. Этому должна предшествовать соответствующая договоренность, достигнутая по открытой связи. В соответствии с данным запросом ЦСФРК обеспечивает генерацию и распределение открытых ключей между пользователями.

Формирование открытых ключей основано на применении описанной однонаправленной функции, использующей логический вывод на перестановках. В ЦСФРК хранятся все ключи-перестановки столбцов и строк, позволяющие из таблицы главного ключа сформировать для каждого пользователя таблицы начального, внутреннего и внешнего секретных ключей. После загрузки системы все эти таблицы, включая таблицу внешних секретных ключей, для разных пользователей будут асимметричны. С целью организации закрытой связи между пользователями А и В необходимо привести их таблицы внешних секретных ключей в симметричное состояние. Это достигается благодаря наличию в ЦСФРК всех указанных функционально связанных секретных перестановок таблиц. При этом с помощью логического вывода на последовательности транзитивной связи между строками таблиц секретных перестановок определяются относительные перестановки для пользователей А и В, которые позволяют привести таблицы внешних секретных ключей в

идентичное состояние. Указанные относительные перестановки являются открытыми ключами; с их помощью пользователи могут перевести таблицы внешних секретных ключей в идентичное состояние для организации симметричной закрытой связи.

Отметим, что функция формирования открытых ключей с использованием относительной перестановки является однонаправленной для любого пользователя системы. На основе полученных открытых ключей в пользовательском устройстве А и сервере В распределенной обработки создают таблицы симметричных внешних секретных ключей. Эти таблицы записываются в одноразовый многоалфавитный кодер (декодер) с целью установления закрытой симметричной связи между пользователями. При этом в процессе шифрования на основе генерации случайных перестановок таблиц внешних секретных ключей реализуется описанный режим одноразовой системы с открытой передачей ключей, который обеспечивает требуемый гарантированный уровень защиты информации. После завершения сеанса закрытой связи ЦСФРК посылает пользователям А и В открытые ключи перестановки для генерации новых асимметричных таблиц внешних секретных ключей.

Предложенная в [5] система обеспечивает возможность эффективного, с гарантированной надежностью обмена зашифрованной информацией. Каждый сертифицированный пользователь, обратившись к ЦСФРК, сможет обмениваться закрытой информацией с любым сервером или пользователем компьютерной сети.

Литература

1. Молдовян А.А. и др. Криптография: скоростные шифры. СПб.: БХВ-Петербург, 2002.
2. Романец Ю.В., Тимофеев П.А., Шаньгин В.Ф. Защита информации в компьютерных системах и сетях. // М.: Радио и связь, 1999.
3. Введение в криптографию / Под общ.ред. В.В. Ященко. // М.: МЦНМО: "ЧеРо", 1999.
4. Насыпный В.В. Комплексная защита процесса обработки информации в компьютерных системах от несанкционированного доступа, программных закладок и вирусов. // М.: МГГУ, 2000.
5. Насыпный В.В. Способ комплексной защиты процесса обработки информации в компьютерных системах и система для осуществления способа. Международная заявка на изобретение № PCT/RU 01/00272 от 05.07.2001.

10. Система с абсолютной стойкостью

Несмотря на то, что почти все применяемые сегодня шифры могут быть раскрыты [1], существует абсолютно стойкий шифр – одноразовая система шифрования, предложенная Г. Вернамом еще в 1926 году [2]. Для ее

реализации применяется одноразовый блокнот, который состоит из нескольких таблиц со случайными числами (ключами). Блокнот имеется в двух экземплярах: один — для отправителя, другой — для получателя. Для каждого символа сообщения применяется свой ключ из таблицы, причем только один раз — после передачи сообщения таблица уничтожается, а шифрование нового осуществляется с помощью новой таблицы. Ясно, что количество одноразовых ключей у каждой пары абонентов должно быть достаточным для передачи всех сообщений и такой шифр абсолютно надежен, если набор ключей таблицы случаен и непредсказуем.

Теоретически доказано, что одноразовые системы шифрования дешифровать невозможно, поскольку зашифрованный текст не содержит достаточной информации для восстановления открытого текста [2], однако одноразовые системы шифрования применяются для связи только между двумя абонентами. В сетях, где необходима зашифрованная связь между всеми абонентами, практически одноразовые системы использовать невозможно — для этого понадобилось бы хранить множество одноразовых блокнотов и по мере применения заменять эти блокноты у каждой пары пользователей.

В сетевой одноразовой системе, где каждый абонент имеет только один комплект асимметричных одноразовых ключей, можно установить зашифрованную связь между абонентами с помощью процедуры перекодирования. Данная процедура заключается в замене ключевой последовательности, используемой для шифрования сообщения отправителя на ключевую последовательность получателя без раскрытия содержания зашифрованного сообщения [3], благодаря чему сохраняется абсолютная стойкость шифра. В состав системы входят Центр сертификации, формирования и распределения ключей (ЦСФРК), серверы распределенной обработки и пользовательские устройства (абонентские комплекты) [3]. Через ЦСФРК осуществляется подключение абонентских комплектов к системе защиты, их сертификация, формирование и распределение ключей, а также организация засекреченной связи между любыми абонентами сети в одноразовом режиме с применением перекодеров. Последние обеспечивают перевод информации, зашифрованной одноразовым ключом одного абонента, в информацию, зашифрованную ключом другого абонента.

В ЦСФРК с использованием датчика случайных чисел для каждого из M абонентов сети формируется заданное множество N одноразовых, случайных и независимых ключей, каждый из которых представляет собой таблицу размером n*n (где n — число символов), заполненную кодами длиной m. Полученные таблицы являются асимметричными одноразовыми ключами (ключевыми таблицами) для каждого из M абонентов сети. Вместе с тем для каждого абонента формируются одноразовые сетевые блокноты, каждый из которых содержит N ключевых таблиц. Каждой произвольной ключевой таблице с номером i присваивается уникальный стохастический индекс Iξi(k), формируемый с помощью специальной хеш-функции, который однозначно определяет данную таблицу. В результате получают одноразовый

сетевой блокнот для каждого из M абонентов, содержащий N одноразовых ключевых таблиц. При необходимости этот ряд ключей засекречивается с помощью шифратора, например, стохастического кодера [4]. Таким образом в ЦСФРК образуется M * N одноразовых ключевых таблиц. Полученные блокноты из N одноразовых ключевых таблиц записываются в M комплектов флэш-памяти — для каждого из M абонентов сети. Флэш-память со множеством одноразовых (зашифрованных) сетевых блокнотов выдается очередному абоненту при его сертификации в ЦСФРК вместе со смарт-картой, содержащей пароль и PIN-код данного пользователя. Полученные флэш-память и смарт-карта устанавливаются в абонентские комплекты каждого абонента для организации сетевой засекреченной связи. Таким образом, для организации сетевой связи между M абонентами используются лишь K=M одноразовых блокнотов.

Именно благодаря перекодированию в ЦСФРК обеспечивается засекреченная связь абонента с любым из Mi абонентов, хотя их одноразовые ключевые таблицы случайны, независимы и асимметричны. Указанная функция реализуется в ЦСФРК при обращении к нему пары абонентов для организации зашифрованной связи. В общем случае ключевые таблицы записываются во флэш-память и применяются в исходном незашифрованном виде. Перед началом сеанса связи с помощью ЦСФРК выполняются идентификация и аутентификация абонентов, например, с использованием алгоритмов, приведенных в работе [3]. Затем происходит выборка из флэш-памяти комплектов абонентов зашифрованных одноразовых блокнотов, которые указаны в значениях хеш-функций Iξi(k), Iξi(k). Эти значения передаются из ЦСФРК абонентам и записываются в блок управления каждого из них. Если заявлен продолжительный сеанс, то ЦСФРК передает столько значений хеш-функций ключевых таблиц различных карт, сколько потребуется для засекречивания в одноразовом режиме всего сеанса связи. Далее по специальной команде ЦСФРК происходит поочередная расшифровка данных таблиц в абонентских комплектах. Если таблицы во флэш-памяти не зашифрованы, то они применяются в исходном виде.

За счет создания соответствующей схемы перекодеров ЦСФРК возможна также одновременная связь одного абонента с заданным множеством других пользователей.

На основе перекодеров ЦСФРК, представленных в виде «коммутаторов» шифров, можно создавать сложные сетевые системы шифрованной связи. Каждый узел коммутации, соединенный с другими перекодерами сети, обеспечивает связь абонентов своей подсети с абонентами любого другого узла коммутации с использованием одноразовых ключей. При этом между любыми абонентами подсети поддерживается связь с абсолютной стойкостью шифрования с применением нескольких перекодеров — по числу задействованных в организации связей в ЦСФРК. Именно такая система легла в основу тестовой сети засекреченной мобильной связи.

Важнейшей областью применения системы с абсолютной стойкостью шифрования могут быть компьютерные стохастические системы [4], обеспечивающие комплексную защиту компьютеров и сетей от программных закладок и вирусов с гарантированным уровнем стойкости. Для достижения такого уровня стойкости в каждый компьютер вводится локальный ЦСФРК, взаимодействующий с системным ЦСФРК для обеспечения функций шифрования передачи, хранения и обработки программ и данных в зашифрованном виде с абсолютной стойкостью. При этом системный и локальный ЦСФРК решают задачи взаимодействия зашифрованных различными одноразовыми ключами программ и данных без раскрытия их содержания, поэтому выполнение программ, информационно-логическая обработка данных и арифметические вычисления могут быть реализованы в зашифрованном виде с абсолютной стойкостью в созданном контуре защиты. Этот контур не имеет точек, уязвимых для информационных атак хакеров, программных закладок и вирусов.

В настоящее время ведутся работы над использованием системы в робототехнике с целью создания «интеллектуальных» роботизированных установок, «общающихся» на естественном языке с абсолютным уровнем защищенности.

Литература

1. Романец Ю.В., Тимофеев П.А., Шаньгин В.Ф. Защита информации в компьютерных системах и сетях. Под ред. В. Ф. Шаньгина. - М.: Радио и связь, 1999.
2. Введение в криптографию. Под ред. В.В. Ященко. - М.: «ЧеРо», 1999.
3. Насыпный В.В. Способ комплексной защиты процесса обработки информации в компьютерных системах и система для осуществления этого способа. Международная заявка на изобретение № PCT/RU 01/00272 от 05.07.2001.
4. Владимир Насыпный, Защищенные стохастические системы. «Открытые системы», 2004, № 3.

11. Защита поиска

Зашифрованные интеллектуальные поисковые системы

Современные поисковые системы становятся все более интеллектуальными, что, однако, не должно ослаблять безопасности доступа к информации. Использование стохастической информационной технологии [1] позволяет комплексно повысить «интеллект» поисковой системы без ослабления ее защищенности. Это стало возможно за счет случайного кодирования и хэширо-

вания символьной информации с целью ее адаптации к конкретной компьютерной среде. В заявке на изобретение № PCT/RU 02/00258 от 28.05.2002 «Способ синтеза самообучающейся системы извлечения знаний из текстовых документов для поисковых систем» было доказано, что имеется возможность создания на основе новой технологии интеллектуальных систем точного поиска, реализующего функции извлечения знаний из текстов и формирования ответов, релевантных запросам пользователей. Вместе с тем, применение стохастической информационной технологии позволяет решить и другую задачу — обеспечение безопасности поиска зашифрованной текстовой информации различного уровня конфиденциальности. При этом создается замкнутый безопасный поисковый контур. Запрос, поступивший от пользователя, шифруется и передается в поисковую машину, где, не расшифровываясь, подвергается дополнительному шифрованию. Это обеспечивает реализацию процедуры интеллектуального поиска на зашифрованных текстовых документах, без раскрытия их содержания. Полученный ответ, релевантный запросу, также будет зашифрован, передан по линии связи и расшифрован на рабочем месте пользователя. Таким образом, исключается возможность доступа к информации, хранящейся в текстовых документах поисковой системы, а также доступ к содержанию вопросов и ответов, передаваемых по сети. Это открывает новые возможности в области создания безопасных поисковых систем, работающих с конфиденциальной информацией.

Применение стохастической информационной технологии позволяет комплексно решать проблему реализации точного поиска и обеспечения безопасности информации. Под точным поиском понимается нахождение системой ответа, релевантного запросу пользователя. При этом запрос формулируется на естественном языке в виде вопросительного предложения. Точный поиск предполагает нахождение ответа с максимально возможной релевантностью — мерой, определяющей, насколько полно тот или иной документ отвечает критериям, указанным в запросе. Точный поиск может быть получен в виде одного предложения текста (краткий ответ) или группы предложений (подробный ответ). При этом критерием релевантности является возможность эквивалентного преобразования с помощью интеллектуальной обработки полученного ответа к виду запроса. Если такое преобразование возможно, то полученный ответ считается в полной мере релевантным запросу или точным. В противном случае производится попытка повторного формирования ответа с использованием дополнительной текстовой информации. Если получение указанного ответа на предоставленном объеме текстовой информации невозможно, то считается, что в данном случае точный ответ не может быть получен.

Точный ответ либо непосредственно содержится в текстовой информации в виде одного или нескольких предложений, либо на основе имеющейся информации происходит извлечение знаний из документов и формируются новые предложения, релевантные запросу, которых в явном виде в тексте нет. Важнейшую роль в этом процессе играют семантический анализ текстовой

информации и логическая обработка фрагментов текста с целью получения новых, семантически связанных текстовых структур, соответствующих требованиям точного ответа.

Основные принципы построения и функционирования системы точного поиска на основе стохастической информационной технологии описаны в [1], а в данной статье мы более детально опишем реализацию семантического анализа и логической обработки текстовой информации в зашифрованном виде с целью формирования точного ответа.

В общем случае зашифрованная система точного поиска включает базу зашифрованных текстовых документов и криптографически защищенные средства ее интеллектуальной обработки: стохастически индексированные базы знаний грамматического и семантического анализа, базы знаний, определяющие правила эквивалентного преобразования, подсистему логического вывода и библиотеку прикладных зашифрованных программ, непосредственно реализующие функции поиска и обработки стохастически преобразованной информации. Выполнение программ также осуществляется в зашифрованном виде, что в сочетании с зашифрованной обработкой информации создает комплексную защиту системы от хакеров, программных закладок и вирусов.

При формировании базы текстовых документов поисковой системы производится стохастическое кодирование символьной информации. Стохастическое индексирование выполняется с использованием специальной хэш-функции, которая обеспечит преобразование различных элементов текстовой информации в их хэш-значения, представленные в виде двоичной комбинации заданной длины, которые принимаются в качестве стохастических индексов. За счет свойств хэш-функции и выбора длины комбинации индекса достигается их гарантированная уникальность для различных элементов текста со сколь угодно малой заданной вероятностью коллизий [1]. При этом сначала формируются стохастические индексы отдельных слов (их основ), которые затем используются для получения индексов словосочетаний, входящих в предложения текста, и самих предложений. На основе стохастических индексов предложений получают индексы абзацев. Названия глав, разделов и самих текстовых документов также преобразуют в соответствующие стохастические индексы.

Полученные индексы обеспечивают произвольный доступ к соответствующим элементам и структурам текстовой информации, которые при этом стохастически кодируются с использованием одноразовой системы шифрования с открытой передачей ключей. Ключи, применяемые при шифровании текстов, записываются в конце каждого зашифрованного предложения. Для перевода слов или словосочетаний из одной системы шифрования в другую используются процессы перекодирования символьной информации без раскрытия ее содержания. Для раскодирования текстовой информации имеются соответствующие декодеры. При этом система формирования и передачи одноразовых открытых ключей обеспечивает реализацию в реальном времени описанных функций кодирования, перекодирования и декодирования тексто-

вой информации. Отметим, что после каждого обращения к соответствующему массиву зашифрованного текста происходит его перешифровка с использованием нового открытого ключа.

В предложенной системе стохастической индексации формирование индексов непосредственно на основе самих символьных объектов обеспечивает возможность ввода новых, исключения старых объектов, изменения порядка их следования, а также модификацию сетевых структур баз знаний в реальном масштабе времени. При этом происходит автоматическая модификация только тех структур, которые непосредственно связаны с вновь вводимыми или исключаемыми объектами, без изменений всей индексной системы. В этом принципиальное отличие стохастического индексирования от регулярного индексирования текстовых документов, при котором любое изменение состава символьных объектов или их связей требует полной реструктуризации системы. Полученная стохастическая индексная система является открытой к изменению состава и содержания поисковой системы в процессе ее функционирования, что делает возможным применение широкой адаптации индексирования к процессам поиска для повышения скорости обработки при проведении семантического анализа текстов. Например, в ходе анализа часто возникает необходимость поиска соответствующих фрагментов текста не только по отдельным словам, но и по словосочетаниям, определяющим различные термины, понятия, предикативную основу, а также другие типы отношений в предложении. Для этого в системе реализована возможность быстрого перехода от индексов отдельных слов к индексам указанных словосочетаний. В результате, обеспечивается произвольный доступ к текстовой информации с целью нахождения нужных предложений, а также выполнение функций логического вывода, классификации и рубрикации текстов. Индексные таблицы автоматически модифицируются для включения строк, связывающих индексы отмеченных словосочетаний с индексами соответствующих предложений абзацев и текстов. За счет этого повышается скорость реализации семантического анализа.

Все перечисленное относится также к построению и функционированию баз знаний, основанных на стохастически индексированных правилах продукций. Применение стохастических индексов предикатов, процедур и правил позволяет образовывать сетевые структуры, в которых время логического вывода линейно зависит от числа используемых правил продукций. При этом полностью снимается проблема «комбинаторного взрыва», характерного для существующих продукционных систем, и обеспечивается реальное время логической обработки независимо от объема базы знаний. Образованная сетевая структура правил продукций является открытой к изменению их состава и содержания. Часто используемые цепочки правил могут быть преобразованы в одно правило путем их агрегации, что повышает скорость обработки информации текстов при семантическом анализе и поиске. Отметим, что построение правил продукций на основе стохастических индексов априори шифрует содержание правил и логику их обработки.

Как известно, цель семантического анализа — анализ смысла составных частей каждого предложения. Для этого в описываемой интеллектуальной поисковой системе используется процесс извлечения знаний из лингвистической литературы. Применяются стохастически индексированные толковые и семантические словари, проблемно-ориентированные словари терминов и определений, энциклопедии, справочники, учебные пособия и др. За счет этого реализуется режим самообучения поисковой системы с использованием логического вывода в указанных текстах, с автоматическим накоплением знаний для проведения грамматического и семантического анализа. Сформированные базы знаний содержат как процедурные знания в виде правил продукций, так и семантические сети, включающие термины и наименования объектов предметной области, предикативные основы предложений текста, а также словосочетания, описывающие типы отношений в каждом предложении.

Запрос, обработка текста, ответ Для иллюстрации представим запрос, сформированный пользователем, предварительно выбранный абзац в процессе анализа текста и полученный точный ответ в открытом виде.

ЗАПРОС: Какие устройства персонального компьютера называются периферийными?

ПРЕДВАРИТЕЛЬНО ВЫБРАННЫЙ АБЗАЦ: Персональный компьютер предназначен для создания, хранения, обработки и передачи данных. Он состоит из различных блоков и устройств. При этом устройства, расположенные внутри системного блока, называются внутренними. Устройства, расположенные снаружи — внешними. Дополнительные подключаемые внешние устройства относятся к периферийным устройствам. Принтер для печати информации на бумаге — пример периферийного устройства.

ТОЧНЫЙ ОТВЕТ: Дополнительные подключаемые внешние устройства персонального компьютера (например, принтер для печати информации) называются периферийными.

В процессе формирования точного ответа в качестве базового было выбрано следующее предложение: «Дополнительные подключаемые внешние устройства относятся к периферийным устройствам». Затем, используя отношения «часть — целое», в него было введено словосочетание «персональный компьютер» из первого предложения абзаца в соответствующем падеже (внешние устройства — часть компьютера). После этого, применяя отношения «род — вид», в базовое предложение включено словосочетание «принтер для печати информации» из последнего предложения абзаца (принтер для печати информации относится к классу периферийных устройств). К этому словосочетанию было добавлено вводное слово «например». Полученная группа слов «(например, принтер для печати информации)» представлена в базовом предложении как вставная конструкция и, соответственно, выделена скобками. Словосочетание «относятся к периферийным устройствам» заменяется на близкое по смыслу словосочетание «называются периферийными устройствами». В итоге выполненного семантического анализа и логической обработки

текста формируется точный ответ.

Для контроля релевантности полученного ответа преобразуем его к виду запроса. При этом группа слов определения «Дополнительные подключаемые внешние» заменена на вопросительное слово «какие». Также была исключена вставная конструкция, которая имеет уточняющее значение. В результате из сформированного ответа получено вопросительное предложение «Какие устройства персонального компьютера называются периферийными?», которое идентично запросу. Это доказывает релевантность полученного точного ответа запросу пользователя.

Для шифрования используется одноразовый многоалфавитный кодер. Зашифрованная информация будет иметь следующий вид (с некоторыми сокращениями ввиду иллюстративного значения данного примера):

ЗАПРОС: ЖЛIUЪЗЬNLQTПWKHEMRAФЮFOЗPTGDSAЩГBWFZИ БОДСESVHВРЙIYKУИГДFDZKРЮФДTVQS

ПРЕДВАРИТЕЛЬНО ВЫБРАННЫЙ АБЗАЦ: WFRCKTЩГOQNЩИ-БSDVИВЙЬVGYЯXMЪYDЖЕГFOЯЧ MЦZEKДВЩЮZHYЖТГ-БIQFBXШЧHWSHЪЖИДБVZRKH ВЮMXTCФЯУKWQЪСЮАЦWLUЩСТ-KQMSHФЗЬЦLREWЖ ФНЮАЩDZXOРЙЧZYВНЯБATNWMЪФГ ЙFQJК-НЮШOG СЖЭВЫVPSЪДЬAKP

ТОЧНЫЙ ОТВЕТ: КРАЦАIFДБMTHJMВЛГЙСКОЕНЩФDBPLУРЗЪ-UWIЖ ЫШTRGЯЬИЭZSЮXOCNQVЧАЕЛЮЦЕКНYXOЗБШSMOGMТЯ Именно в таком виде информация может попасть к злоумышленнику при попытке несанкционированного доступа к системе.

Отметим, что пользователь имеет доступ к содержанию запроса (до его кодирования с целью передачи в поисковую систему), а также полученного точного ответа (после его декодирования). Вся текстовая база системы, включая приведенный предварительно выбранный абзац, является для пользователя зашифрованной.

Декодирование ответа осуществляется в компьютере пользователя, выдавшего исходный запрос. При этом передача по линии связи также осуществляется в зашифрованном виде после необходимого перекодирования зашифрованного вопроса или ответа.

***

В предлагаемой системе реализуется полностью замкнутый зашифрованный контур точного поиска информации с выполнением функций интеллектуальной обработки текстов, включающих необходимые элементы семантического анализа. Если смотреть на эту проблему шире, то создание подобных систем гарантирует любому пользователю полную конфиденциальность диалога. Это, в рамках существующих законов, обеспечивает права каждого на защиту его личной информации.

12. Проекты

Инновационные технологии и проекты

В настоящее время группой российских ученых и исследователей под руководством Владимира Владимировича Насыпного (профессор, доктор технических наук, известный ученый, автор ряда патентов, имеет опыт разработки научного базиса инновационных международных IT проектов) создан, запатентован и проверен на практике целый комплекс абсолютно инновационных информационных технологий. Данные технологии базируются на фундаментальных разработках и ориентированы на создание проектов, как инструментальных средств, так и прикладных программных комплексов и систем.

В основу указанных технологий положены не имеющие аналогов в мире принципиально новые научно-технические решения в области развития информационных систем и обеспечения их безопасности, которые, во-первых, позволяют создать и вывести на рынок абсолютно новый класс технических средств, обладающих реальным искусственным интеллектом, во-вторых, позволяют существенным образом повысить эффективность систем контроля, управления и принятия решений любого уровня, в том числе, нацеленных на решение задач национальной безопасности, и, в-третьих, могут стать базовыми для прорыва страны (IT-компании) в мировые лидеры IT-технологий.

Применение стохастики позволит в кратчайшие сроки (2-3 года) пройти путь от теории и практики больших данных к индустрии знаний и нанотехнологии-квантовым компьютерам с интеллектом. Это даст возможность разработать принципиально новые технические системы практически во всех промышленных сферах, включая новые поколения компьютеров, суперинтеллектуальные защищенные системы в робототехнике, в авиакосмической, атомной промышленности и других высокотехнологичных областях. Могут быть созданы новые поколения интеллектуальных транспортных средств, бытовой техники, «умных» вещей и других востребованных на мировом рынке товаровмассового спроса.

В результате стохастика может обеспечить эффективное решение задачи базовой модернизации промышленности и создания миллионов рабочих мест в ближайшие пять лет.

К первоочередным разработкам, прежде всего, следует отнести технологии и проекты создания систем:

- автоматического распознавания и понимания смысла слитной речи от неизвестного диктора с неограниченным объемом словаря, обеспечивающих эффективную защиту от внешних помех и артикуляционных искажений речевого сигнала;
- автоматического семантического анализа, поиска по запросам пользователей, обработки и извлечения из неструктурированной текстовой информации знаний, релевантных этим запросам;

- аналитической обработки больших объемов неструктурированной текстовой информации, обеспечивающих возможность реализации режима самообучения с извлечением знаний из текстов для проведения глубинного семантического анализа;
- проактивной защиты от деструктивных опасных программ и несанкционированного доступа, обеспечивающих создание нового поколения безопасных компьютеров, выполняющих процессы передачи, хранения, обработки данных и выполнения программ в зашифрованном виде с гарантированным исключением точек доступа к содержанию вычислительного процесса со стороны вредоносных воздействий и хакеров;
- распознавание и синтеза сложных изображений трехмерной графики предельной информационной емкости, их символьной интерпретации, семантически связанной с полученными визуальными образами в реальном времени, обеспечивающих смысловую оценку ситуаций в зоне наблюдения при комплексном использовании звукового и визуального каналов контроля.

В этом ряду следует выделить принципиально новую технологию создания систем распознавания и понимания смысла речи.

Исследования по данной проблеме продолжаются крупнейшими мировыми компаниями и фирмами в течение более, чем 50 лет. При этом до настоящего времени эффективных решений не найдено и универсальных прикладных систем распознавания речи неизвестного диктора, использующего неограниченный словарь, никем не построено.

Учеными группы В.В. Насыпного доказано

1) известные научно-методические подходы к решению задач распознавания естественной речи в реальных шумовых условиях, опирающиеся на создание акусто-фонетических баз различных дикторов и формирование на этой основе некоего «усредненного» эталонного диктора, достигли своего теоретического предела и не позволяют кардинально повысить качество распознавания речи в условиях появления нового диктора;

2) преодолеть данный предел возможно на основе выявления и использования механизмов связи процессов распознавания речи с пониманием ее смысла путем семантического анализа непосредственно речевых образов, обрабатываемых в системе; при этом реализуются два параллельных процесса обработки речевой информации: в образном виде (обработка звуковых образов, описывающая их семантически) и логическая обработка символьной информации; необходимо отметить, что этот подход является аналогом двуединства образного и рационального мышления человека.

Предлагаемый подход положен в основу разработанных новых методологии и технологии распознавания слитной речи неизвестного диктора, которые базируются на создании самообучающихся систем извлечения знаний из символьных описаний речевых образов, а также на их семантическом анализе. Предлагаемые решения запатентованы в России и за рубежом, а их экспериментальная проверка показала практически 100%-ую

достоверность распознавания даже при наличии сильных внешних помех и артикуляционных искажений речевого сигнала. Тем самым, создан инновационный научно-технический базис для построения прикладных систем чрезвычайно широкого спектра.

Одновременно с этим разработана и запатентована принципиально новая технология создания высокоэффективных автоматизированных информационно-аналитических систем. Все существующие в настоящее время технологии и информационно-аналитические системы, работающие на массивах неструктурированной текстовой информации, характеризуются высокой долей «ручных» операций как при настройке подобных систем на прикладную область, так и в процессе собственно аналитической работы.

С целью повышения уровня автоматизации информационно-аналитической деятельности разработанная специалистами группы В.В.Насыпного технология позволяет на основе полного лингвистического анализа неструктурированных текстов реализовать автоматизированный разбор, обобщение и синтез семантической информации и обеспечивает извлечение знаний, релевантных аналитическим запросам пользователей, типа: Почему…?, Чем вызвано…? Что делать, если…? и т.п.

Задача аналитической обработки текстовой информации решается в общем контуре создания самообучающихся систем с извлечением знаний из текстов, их семантического анализа и логического вывода. Это впервые позволяет решить проблему автоматического формирования баз знаний по заданным предметным областям и их эффективной аналитической обработки с использованием функций анализа, синтеза, дедукции, индукции, определения, обобщения, сравнения и аналогии. При этом использование разработанной стохастической информационной технологии обеспечивает возможность логического вывода на практически неограниченном объеме исходной информации с полным устранением эффекта «комбинаторного взрыва», который является неразрешимой проблемой для традиционных систем обработки текстов.

Предлагаемая методология создает условия для аналитической обработки информации в реальном масштабе времени и в рамках любых заданных предметных областей (политика, социология, наука, технологии, экономика, государственная безопасность и другие). При этом обеспечивается существенное (в разы) повышение полноты охвата информации, результативности, достоверности и оперативности аналитической обработки, что практически недостижимо для существующих аналитических технологий с участием коллективов экспертов.

Специалистами группы В.В. Насыпного также разработан, запатентован и реализован на уровне прототипа ряд инновационных технологий в области информационной безопасности.

В настоящее время в существующих компьютерных системах не решена проблема гарантированной защиты баз данных, содержащих информацию ограниченного доступа. Это прежде всего относится к компьютерным системам государственных учреждений. Данная проблема

существует также в информационно-управляющих системах, обеспечивающих функционирование сетей электро- и водоснабжения, транспортной сети, телекоммуникационных систем, учреждений финансовой и банковской сферы, а также в других системах с высокими требованиями к уровню защиты информации. Особенно это важно для реляционных баз данных типа Oracle, Microsoft SQL и других систем, получивших широкое распространение. Сюда же можно отнести персональные компьютеры, базы данных которых содержат информацию личного характера.

В настоящее время надежная защита информации в базах данных осуществляется только при их хранении на магнитных носителях. При необходимости поиска требуемой информации в базах данных по запросам пользователя происходит расшифрование файлов информации и обработка ее в компьютерах в открытом виде. Кроме того, при поступлении запросов в зашифрованном виде по каналам связи они расшифровываются в интерфейсах подключения линий связи к компьютеру и вводятся в него в открытом виде. В это время обрабатываемая информация является открытой для несанкционированного доступа, атак хакеров, компьютерных вирусов и программных закладок.

Таким образом, в существующих компьютерных системах не обеспечивается единый гарантированный контур защиты распределенной обработки информации, включающий защиту как обработки и хранения информации в базах данных, так и передачи по каналам связи.

Решение указанной задачи разработано В.В. Насыпным на основе стохастической информационной технологии и базируется на введении стохастичности в вычислительный процесс с использованием одноразовых систем шифрования информации и специальных функций хэширования данных при их обработке, хранении и передаче. За счет введения нового функционального элемента - перекодера, который дает возможность перешифровывать информацию, не раскрывая ее содержания, образуется сквозной контур защиты от несанкционированного доступа. При этом интеллектуальные возможности стохастической технологии позволяют производить поиск и обработку информации в зашифрованной базе данных без раскрытия ее содержания. При поступлении запроса пользователя обработка данных может производиться в зашифрованном виде на основе единого метода логического вывода с использованием стохастических индексов, отображающих концептуальное описание содержимого реляционных таблиц.

Для реализации защиты баз данных в сервере создается специальный криптографический универсальный программный комплекс (УПК). Этот комплекс выполняет функции формирования одноразовых ключей, хэширования информации, получения стохастических индексов, перекодирования информации, а также формирует концептуальное описание реляционных баз данных в виде древовидных логических структур и обеспечивает их обработку с целью поиска в базе требуемых зашифрованных

данных. Логический вывод по древовидным структурам позволяет найти не только ячейки таблицы, содержащие необходимые, указанные в запросе элементы данных, а также определить соответствующие им одноразовые ключи, которые затем используются в перекодере для их перешифрования с целью выдачи зашифрованного ответа по каналу связи пользователю, пославшему запрос. При этом содержание данных не раскрывается. Расшифрование полученных данных производится в декодере пользователя с использованием идентичного одноразового ключа. При необходимости реализовать в ходе поиска в реляционной базе логические функции производится логический вывод с использованием стохастических индексов ключевых элементов реляционных таблиц. Для проведения арифметических вычислений с зашифрованными элементами данных применяется специальная программа, которая впервые обеспечивает реализацию необходимых вычислений в защищенном, стохастически преобразованном виде. Кроме этого, в криптографическом программном комплексе происходит аутентификация, идентификация и контроль целостности программ и данных.

Таким образом, разрабатываемый универсальный программный комплекс на основе стохастической технологии обеспечивает гарантированный замкнутый контур процессов передачи, обработки и хранения информации в реляционных базах данных. Аналогов подобному комплексу среди разрабатываемых систем защит в мире не существует. Способ и система, положенные в основу УПК, защищены российскими и зарубежными патентами.

Разработанные учеными группы В.В. Насыпного технологии и прототипы продукции позволяют создавать сложные интеллектуальные системы, по своим показателям существенно превышающие все наиболее конкурентоспособные мировые аналоги.

При определенной финансовой поддержке группа ученых и исследователей под руководством В.В. Насыпного готова на основе упомянутого научно-технического базиса создать государственно-коммерческую специализированную фирму для разработки целой линейки абсолютно инновационных инструментальных средств и прикладных информационных систем. При этом будет использоваться опыт создания и функционирования фирмы «Стокона», которая создавалась для реализации стохастической информационной технологии при участии международных инвестиционных фондов. В результате была разработана и установлена в США (2005 год) первая в мире интеллектуальная поисковая система NearU.

Готовы к обсуждению любых схем и форм сотрудничества с заинтересованным инвестором. У нас разработаны бизнес-планы для всех указанных выше проектов.

Характеристики рынка создаваемых продуктов/технологий

Спрос на технологии и системы распознавания и понимания аудио и видеоинформации, семантической обработки знаний составляет более $105 млрд. Потенциальный объем международного рынка систем распознавания речи в 2011 году составляет ~ $40 млрд.

Конкуренция на нем невысокая ввиду отсутствия технологий и систем, обеспечивающих достижение характеристик, необходимых для создания коммерческих продуктов.

Системы качественного распознавания речи, смыслового анализа неструктурированной информации и обработки знаний с заявленными в проекте характеристиками на IT рынке отсутствуют.

В перспективе потенциальный объем мирового рынка систем распознавания речи составит более $50 млрд. за счет расширения прикладной области применения технологии и систем.

Потребителями разрабатываемой в проекте технологии и систем являются практически все вертикали рынка IT продукции и услуг: государственные организации, предприятия различных форм собственности, множество независимых производителей или небольшое количество крупных и т.п. Ввиду наличия широкого спектра продукции и услуг, реализующих функции распознавания речевой информации, создаваемая продукция будет востребована в различных уровнях перечисленных отраслей (например, на уровнях управления отраслью, производственных предприятий, услуг отраслевых предприятий и т.п.).

Аналогов разрабатываемой технологии и системам распознавания и синтеза речи, понимания и перевода информации, семантической обработки знаний, проактивной защиты компьютерных систем на мировом IT рынке в настоящее время нет. Используемые в существующих системах методологии и технологии не дают ощутимых результатов, достаточных для создания систем коммерческого применения. Поэтому ниша систем распознавания и синтеза речи, понимания и перевода информации, семантической обработки знаний, проактивной защиты компьютерных систем на мировом IT рынке остается свободной.

Прямых конкурентов в настоящее время нет. Наиболее близкими продуктами являются сервисы обработки информации корпораций Microsoft и Google. Выпуск аналогичной продукции сторонними компаниями невозможен ввиду патентования основных технологических решений, положенных в основу создаваемых технологии и систем.

Характер спроса – равномерный, что обусловлено постоянной потребностью IT рынка в продуктах и технологиях распознавания смыслового содержания аудио и видеоинформации, семантической обработки знаний, защиты компьютерных систем.

Текущий статус проектов

Состояние работ:

- разработаны базовые методы распознавания и понимания слитной речи, эффективность которых подтверждена результатами тестирования на созданном макете;

- разработаны и апробированы базовые алгоритмы и программные модули, доказывающие возможность успешной реализации проекта;

- создан демонстрационный стенд, реализующий распознавание слитной речевой информации для ограниченного словаря системы;

- получен ряд патентов РФ, США, Китая на способы и устройства семантической обработки информации и знаний, свидетельства о государственной регистрации программ для ЭВМ;

- имеются исходные коды программного обеспечения, документация на ранее созданные системы: семантическая поисковая система AskNet (Stocona) Search, эвристический антивирус Stocona Antivirus, система шифрования сети сотовой связи абсолютной стойкости;

- реализован лингвистический процессор автоматического анализа текстов на русском и английском языках программное обеспечение семантической поисковой системы AskNetSearch (www.asknet.ru), занявший 1 место по результатам тестирования РОМИП-2006;

- получены результаты поисковой работы по разработке методов повышения эффективности алгоритмов автоматического распознавания слов в естественной русской речи".

Возможность успешного завершения проекта обусловлена наличием:

- теоретических основ распознавания и понимания смысла речи на основе стохастики в шумах (одноименная монография В.Насыпного);

- имеющихся заделов в виде результатов ранее проведенных поисковых и прикладных работ по созданию систем распознавания речи и обработки неструктурированной текстовой информации;

- возможности привлечения материально-сырьевых и финансовых ресурсов, кадров необходимой квалификации, производственных мощностей и инфраструктуры, необходимых для выпуска опытного образца изделия;

- соответствующих патентов и свидетельств на программы обработки неструктурированной текстовой информации;

- прототипов программного обеспечения систем, доказывающих принципиальную реализуемость предлагаемых к разработке технологий.

На базе прототипов глобальной модификации семантической поисковой системы Заявителями разработан и в настоящее время функционирует в Интернете ряд информационно-поисковых порталов, реализующих семантический поиск информации. Данные порталы предоставляют автоматические вопросно-ответные поисковые сервисы пользователям Интернета.

Разработаны прототипы коммерческих версий корпоративных и персональных семантических поисковых систем.

Поисковая система AskNet (Stocona) Search заняла первое место в тесте дорожки вопросно-ответного поиска семинара РОМИП-2006, существенно опередив разработки РАН (примерно в 2 раза лучше результаты по сравнению

с системой Exactus), получила диплом победителя конкурса лучших программных решений выставки Softool-2004 в номинации «Интеллектуальные поисковые системы», получила серебряную медаль салона инноваций Архимед-2010 и медаль выставки «Высокие технологии 21 века». Наши патенты получили медали на выставках в Брюсселе и в Париже. Проект создания технологии и системы распознавания и понимания речи на любом национальном языке получил Золотую медаль Международного конкурса инновационных проектов на Экспо -2010 в Шанхае.

Аналогичных по технологическому содержанию работ на мировом IT рынке не проводится.

Литература

1. Насыпный В.В. Развитие теории построения открытых систем на основе информационной технологии искусственного интеллекта. М.: Воениздат, 1994. — 248 с.

13. Интеллектуальная система экспертизы проектов

Концепция создания системы

Создание эффективной и конкурентоспособной экономики в современных условиях невозможно без ускоренного развития науки и новых технологий. Прежде всего это относится к информатизации и автоматизации, охватывающей как органы управления различного уровня, так и непосредственно производственные системы. Вступление России в ВТО обуславливает необходимость создания в нашей стране индустрии, способной производить технику и оборудование с мировым уровнем новизны, способной успешно конкурировать на мировом рынке.

Серьезным препятствием этому является несовершенство конкурсного отбора проектов, устаревшая технология работы экспертов, отсутствие индустрии знаний о состоянии развития науки и технике в мире. По данным СМИ за последние десять лет в многочисленные проекты государством было вложено более 300 млрд. рублей. Однако достичь требуемого уровня производства конкурентоспособной научно-технической продукции пока не удалось.

Для устранения указанных препятствий предлагается проект:

«Интеллектуальная система экспертизы проектов, обеспечивающая автоматическое понимание смысла, реферирование, обобщение, определение новизны, реализуемости и эффективности проектных решений».

Данный проект не имеет аналогов в мире. Он разработан на основе стохастики – стохастической информационной технологии.

Как было отмечено в предыдущих разделах, стохастика или стохастическая информационная технология обладает следующими свойствами, без которых невозможно решение проблемы BigData [1], понимание смысла и извлечение знаний из неструктурированных текстов,

распознавание и понимание смысла слитной речи от неизвестного диктора [3], создание эффективных интеллектуальных систем [4, 5,6]:

- реализация логического вывода на больших пространствах поиска с использованием только логически и семантически связанных текстовых структур с исключением перебора на всем пространстве поиска, что обеспечивает исключение комбинаторного взрыва;

- осуществление саморазвития и самообучения системы новым знаниям, определяющим логически и семантически связанные элементы текста, формирование новых знаний, необходимых для получения «картины мира» и семантических классификаторов в различных проблемных областях;

- автоматическое создание баз знаний, описывающих все возможные свойства понятий и логических связей картины мира во всех возможных ситуациях на пространствах поиска объемом не менее 10^{15};

- реализация аналитических и поисковых функций на множестве исходной текстовой информации с использованием полученных знаний в реальном масштабе времени с максимальным пространством поиска не менее 10^{20}

Указанными свойствами обладает только стохастика. Они не доступны для традиционных информационных технологий. Стохастика была разработана в России В.Насыпным и впервые опубликована в монографии «Развитие теории построения открытых систем на основе информационной технологии искусственного интеллекта» (М.: Воениздат, 1994. - 248с.).

Стохастика позволяет создавать суперинтеллектуальные системы, которые могут применяться как для понимания смысла, реферирования, обобщения и определения новизны содержания текстовых документов проекта, так и для моделирования процесса функционирования создаваемых систем, оценки реализуемости и эффективности проектных решений по этапам их разработки. Это создает основу для объективного контроля всего цикла реализации финансируемого проекта и принятия мер по достижению требуемого качества создаваемого изделия, вплоть до своевременной замены фирмы- разработчика.

С целью моделирования процесса функционирования создаваемых систем в предлагаемом проекте предусмотрено создание интеллектуальной информационно-управляющей системы.

Под информационно-управляющими системами, в общем случае, будем понимать проблемно-ориентированные вычислительные сети, которые в дополнении к базовым телекоммуникационным системам имеют интеллектуальную функциональную надстройку и предназначены для комплексного решения задач информатизации и функционирования современной техники. Для моделирования могут применяться локальные, корпоративные и глобальные ИУС.

Внедрение информационной технологии искусственного интеллекта (ИИ) осуществляется на всех уровнях ИУС, от органов управления до исполнительных устройств и оборудования (автоматизированных технологических линий, станков с программным управлением,

промышленных роботов и других средств). Это позволяет говорить о целесообразности создания в рамках ИУС локальных или распределенных информационных и интеллектуальных систем как основы для моделирования процессов функционирования создаваемых технических средств..

Построение распределенных систем невозможно без внедрения телекоммуникационной технологии информационно-вычислительных сетей (ИВС), которые получили широкое распространение во всех видах деятельности, став базой создаваемой индустрии обработки информации.

Это ставит проблему построения интеллектуальных ИУС открытого типа для моделирования взаимодействия распределенных по объектам управления средств обработки данных и знаний с использованием трактов передачи информации ИВС для выработки альтернатив управленческих решений. При моделировании локальных ИУС отдельных устройств применяются различные физические линии связи и интерфейсы.

Особое значение при этом имеет разработка методических и технологических решений обеспечения состоятельности, целостности распределения данных и знаний в процессе их обработки и обновления с использованием ресурсов ИВС или линий связи.

Широкое внедрение в производство гибких технологических линий, постоянное развитие и совершенствование производственных и административных процессов, направленных на реализацию новых видов конкурентоспособной продукции, обуславливают необходимость построения моделирующих ИУС как адаптивных, эволюционно развивающихся систем, открытых к структурному и содержательному изменению предметной области процессов информатизации и автоматизируемых функций.

В комплексе вопросов, связанных с построением интеллектуальных ИУС, определяющим является создание методов обработки символьной информации, позволяющих рационально соединить технологию современных компьютеров, специально не предназначенных для работы со знаниями, с информационной технологией искусственного интеллекта. От этого во многом зависят создание эффективных распределенных и локальных систем обработки данных и знаний и их внедрение в систему управления разрабатываемых изделий.

Настоящий проект посвящен изложению современных подходов к созданию моделирующих ИУС открытого типа. В основу проекта положен принцип системного единства методологического аппарата и информационной технологии систем ИИ и теории телекоммуникационных технологий построения открытых информационных сетей. Этот синтез позволяет, с одной стороны, создавать адаптивные ИУС открытого типа, обеспечивающие эволюционное развитие функциональных возможностей и уровня интеллектуализации моделирующих информационно-управляющих систем. С другой стороны, указанный принцип дает возможность

плодотворного поиска решений в области построения распределенных информационных и интеллектуальных систем, обеспечивающих эффективную обработку, целостность и состоятельность данных и знаний в процессе функционирования ИУС.

Представленные в данном проекте методы, получившие развитие в теории синтеза распределенных систем обработки данных и знаний, основаны на применении стохастической информационной технологии. Она как было отмечено выше, позволяет адаптировать методы представления и обработки знаний к технологии современных компьютеров. Одновременно обеспечивается реализация функций контроля целостности, состоятельности и достоверности информации при ее хранении, обновлении и передачи в системе. В перспективе, разработанная интеллектуальная система позволит реализовать функции автоматического проектирования и синтеза программного обеспечения, разработанных технических средств[4].

Описание проекта будет состоять из введения, пяти глав и заключения.

В первой главе излагаются методы автоматического понимания смысла, обобщения и реферирования неструктурированного текста на основе стохастики. Разработаны методические и технологические основы определения новизны содержания текстовых документов (проектной документации), выявления случаев плагиата.

Во второй главе будут изложены принципы построения моделирующих интеллектуальных ИУС для оценки реализуемости и эффективности проектируемых технических систем различного назначения.. Предложена архитектура функциональной среды (ФС), определяющая предметную область, правила реализации и взаимодействия прикладных процессов интеллектуальных ИУС в ходе моделирования.

Третья глава посвящена обоснованию требований к программному обеспечению интеллектуальных ИУС организационно-административного назначения. В ней представлена проблема разработки метода формализованного описания ИУС. Описаны возможности современных методов представления данных и знаний для описания ФС ИУС. Сформулированы требования к формализованному описанию протоколов и инструментальным системам для синтеза ФС ИУС .

В четвертой главе представлены методические основы создания инструментальной системы для синтеза ФС моделирующих ИУС. В их основу положена стохастическая информационная технология, позволяющая создавать эффективные системы обработки данных и знаний интеллектуальных ИУС.

В пятой главе приведены теоретические основы реализации функций оценки реализуемости и эффективности моделируемых проектных решений. С этой целью разрабатывается система автоматизированного проектирования и управления адаптацией ФС ИУС на основе стохастической информационной технологии. Эта система предназначена для спецификации, контроля корректности и верификации протоколов ФС на этапе проектирования заданного изделия путем адаптации интеллектуальных

систем. В результате производится оценка реализуемости заданных проектных решений. Оценка эффективности осуществляется с использованием разработанных игровых логических моделей.

В заключении обосновывается этапность создания предложенной интеллектуальной системы экспертизы проектов. На первом этапе реализуются функции автоматического понимания смысла, обобщения, реферирования и определения новизны содержания текстов проектной документации. На втором этапе обеспечивается оценка реализуемости и эффективности проектных технических решений. Показано, что создание интеллектуальной системы экспертизы проектов первого этапа обеспечит экономическую эффективность в размере 3-4 млрд. рублей ежегодно.

Литература

1. Черняк Л. Большие Данные — новая теория и практика // Открытые системы №10, 2011.
2. Насыпный В.В. Защищенные стохастические системы // Открытые системы №3, 2004.
3. Насыпный В.В. Распознавание и понимание смысла речи на основе стохастики в шумах. М.: Прометей, 2010. – 139 с.
4. Насыпный В.В. Развитие теории построения открытых систем на основе информационной технологии искусственного интеллекта. М.: Воениздат, 1994. - 248с.
5. Насыпный В.В., Насыпная Г.А. Способ синтеза самообучающейся системы извлечения знаний из текстовых документов для поисковых систем. Патент РФ №2273879, номер международной заявки PCT/RU02/00258, дата подачи 28 мая 2002.
6. Насыпный В.В., Насыпная Г.А. Способ синтеза самообучающейся аналитической вопросно-ответной системы с извлечением знаний из текстов, заявка на патент №2007120344/09 от 06.08.2007. Получено решение на выдачу патента на изобретение от 21.07.2008.
7. Насыпный В.В. Система с абсолютной стойкостью // Открытые системы №9, 2005.
8. Насыпный В.В., Насыпная Г.А. Система распознавания, понимания смысла, анимационного моделирования и синтеза речи на основе стохастической информационной технологии. М.: Прометей, 2008. – 76 с.
9. Искусственный интеллект. Справочник. Кн. 2. Модели и методы. Под ред. Поспелова Д.А. М.: Радио и связь, 1990. - 303 с.
10. Halsall F. Data communications computer networks and osi. Addison-wesley publishing company, 1988. - 973 с.
11. Насыпный В.В. Способ комплексной защиты распределенной обработки информации в компьютерных системах и система для осуществления способа. Патент РФ №2259639, номер международной заявки PCT/RU /00272, дата подачи 28.10.2003г.

12. Насыпный В.В., Насыпная Г.А. Метод семантической связи текста с трехмерной графикой. – М.: Прометей, 2007. – 27с.
13. Кобаяси Н. Введение в нанотехнологию / Н.Кобаяси. – Пер. с японск. – М.: БИНОМ. Лаборатория знаний, 2007. – 134 с.
14. Валиев К.А., Кокин А.А. Квантовые компьютеры: надежды и реальность. – Москва-Ижевск: НИЦ «Регулярная и хаотическая динамика», 2004, 329 стр.

14. Основные научные труды по стохастике

1. Патент РФ №2345416, 31.05.2007 г. "Способ синтеза самообучающейся аналитической вопросно-ответной системы с извлечением знаний из текстов",
2. Патент РФ № 2273879, 28.05.2002 г. "Способ синтеза самообучающейся системы извлечения знаний из текстовых документов для поисковых систем", международная заявка № PCT/RU2002/000258, 28.05.2002 г.,
3. United States Patent Application 20050071150 Nasypny, Vladimir Vladimirovich March 31, 2005 Method for synthesizing a self-learning system for extraction of knowledge from textual documents for use in search,

4. China Patent Application ZL 02 8 29032.1, Method for synthesizing a self-learning system for extraction of knowledge from textual documents for use in search. Priority 04.06.2008,

5. China Patent Application ZL 01 8 23446.1, Method for synthesizing a self-learning system for knowledge acquisition for text-retrieval systems. Priority 11.07.2007,

6. Патент РФ № 2259639, 05.07.2001 г. "Способ комплексной защиты распределенной обработки информации в компьютерных системах и система для осуществления способа ", международная заявка № PCT/RU2001/00272, 05.07.2001 г.,

7. Патент РФ №2137185, 10.09.1999 г. "Способ комплексной защиты процесса обработки информации в ЭВМ от несанкционированного доступа, программных закладок и вирусов",

8. Nasypny V.V. «Method for protecting computer systems against encrypted and polymorphous viruses», PCT/RU2004/000288, WO/2006/022566, 02.03.2006.

9. Насыпный В.В. Развитие теории построения открытых систем на основе информационной технологии искусственного интеллекта. М.: Воениздат, 1994. - 248с.

10. Насыпный В.В. Распознавание и понимание смысла речи на основе стохастики в шумах. М.: Прометей, 2010. – 139с.

11. Насыпный В.В. Распознавание речи на основе интеллектуальных систем. М.: Прометей, 2010. – 59с.

12. Насыпный В.В., Насыпная Г.А. Зашифрованные поисковые системы. Дополнение 1 (с.507-517) к книге Н.Смарта «Криптография» Москва: Техносфера, 2005. - 528с.

13. Насыпный В.В. Сетевая система с абсолютной стойкостью. Дополнение 2 (с.518-523) к книге Н.Смарта «Криптография» Москва: Техносфера, 2005. - 528с.

14. Насыпный В.В., Насыпная Г.А. Способ синтеза самообучающейся аналитической вопросно-ответной системы с извлечением знаний из текстов, заявка на патент №2007120344/09 от 06.08.2007. Получено решение на выдачу патента на изобретение от 21.07.2008.

15. Насыпный В.В., Насыпная Г.А. Метод семантической связи текста с трехмерной графикой. М.: Прометей, 2007. - 27с.

16. Насыпный В.В., Насыпная Г.А. Способ синтеза самообучающейся системы извлечения знаний из текстовых документов для поисковых систем. Патент РФ №2273879, номер международной заявки PCT/RU02/00258, дата подачи 28 мая 2002.

17. Насыпный В.В. Способ комплексной защиты распределенной обработки информации в компьютерных системах и система для осуществления

способа. Патент РФ №2259639, номер международной заявки PCT/RU /00272, дата подачи 28.10.2003г.

18. Насыпный В.В., Насыпная Г.А. Поисковая машина для карманных компьютеров // Мир ПК, 2003, №6, с. 77.

19. Насыпный В.В. Защищённые стохастические системы // Открытые системы, 2004, №8, с. 60-61.

20. Насыпный В.В., Насыпная Г.А. Совершенствование поисковых систем – М.: Прометей, 2006. – 16с.

21. Насыпный В.В., Насыпная Г.А., Огарок А.Л., Тулин В.В. Почему отечественная стохастическая информационная технология устраняет «комбинаторный взрыв» при обработке знаний и позволяет создавать самообучающиеся интеллектуальные поисковые системы // Информатизация и связь, 2009, №1, с. 167-171.

22. Насыпный В.В., Насыпная Г.А. Система распознавания, понимания смысла, анимационного моделирования и синтеза речи на основе стохастической информационной технологии. – М.: Прометей, 2008. – 76с.

23. Насыпный В.В., Насыпная Г.А. Автоматическая классификация понятий в интеллектуальных поисковых системах // Информатизация и связь, 2009, №1, с. 81-85.

24. Насыпный В.В., Насыпная Г.А. Разработка методов и инструментальных средств эволюционного описания предметной области информационно-аналитических систем организационно-административного назначения // Информатизация и связь, 2009, №1, с. 117-124.

25. Насыпный В.В. Стохастика как основа для перехода к большим данным, индустрии знаний и нанотехнологии. – М:МПГУ, 2011. – 24 с.

Содержание

Научное издание

Насыпный Владимир Владимирович

СТОХАСТИКА

Перспективная информационная технология

Подписано в печать 01.09.2012 Формат 60x90/16

Усл.печ.л. 6,75 Тираж 500 экз. Заказ № 88

Отпечатано в типографии МПГУ
129164 г. Москва, ул. Кибальчича, дом 6, стр.2