(12) МЕЖДУНАРОДНАЯ ЗАЯВКА, ОПУБЛИКОВАННАЯ В СООТВЕТСТВИИ С ДОГОВОРОМ О ПАТЕНТНОЙ КООПЕРАЦИИ (PCT)

(19) ВСЕМИРНАЯ ОРГАНИЗАЦИЯ ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ
Международное бюро

(43) Дата международной публикации:
4 декабря 2003 (04.12.2003)

PCT

(10) Номер международной публикации:
WO 03/100659 A1

(51) Международная патентная классификация [7]:
G06F 17/30, G09B 19/00

(21) Номер международной заявки: PCT/RU02/00258

(22) Дата международной подачи:
28 мая 2002 (28.05.2002)

(25) Язык подачи: русский

(26) Язык публикации: русский

(71) Заявители и
(72) Изобретатели: **НАСЫПНЫЙ Владимир Владимирович** [RU/RU]; 115573 Москва, ул. Шипиловская, д. 44/27, кв. 75 (RU) [**NASYPNY, Vladimir Vladimirovich**, Moscow (RU)]. **НАСЫПНАЯ Галина Анатольевна** [RU/RU]; 115573 Москва, ул. Шипиловская, д. 44/27, кв. 75 (RU) [**NASYPNAYA, Galina Anatolievna**, Moscow (RU)].

(74) Агенты: **ЕГОРОВА Галина Борисовна** и др. "Юридическая фирма Городисский и партнеры" ; 129010 Москва, ул. Б.Спасская, д. 25, строение 3 (RU) [**EGOROVA, Galina Borisovna** et al. "Gorodisky & Partners Law Firm Ltd". ; Moscow (RU)].

(81) Указанные государства *(национально)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Указанные государства *(регионально)*: ARIPO патент (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), евразийский патент (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), европейский патент (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), патент OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Опубликована
С отчётом о международном поиске.

В отношении двухбуквенных кодов, кодов языков и других сокращений см. «Пояснения к кодам и сокращениям», публикуемые в начале каждого очередного выпуска Бюллетеня PCT.

(54) **Title:** METHOD FOR SYNTHESISING A SELF-LEARNING SYSTEM FOR KNOWLEDGE ACQUISITION FOR TEXT-RETRIEVAL SYSTEMS

(54) **Название изобретения:** СПОСОБ СИНТЕЗА САМООБУЧАЮЩЕЙСЯ СИСТЕМЫ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ ТЕКСТОВЫХ ДОКУМЕНТОВ

(57) **Abstract:** The invention can be used for developing data-retrieval systems based on the Internet. Said invention makes it possible to automatically form knowledge and to extract said knowledge from electronically presented text-based documents in various languages and to intellectually process text-based data and user requests. The inventive method consists in providing a self-leaning mechanism for a system involving rules of grammatical and semantic analysis in the form of a stochastically indexed intelligence system, forming a database for stochastically indexed dictionaries and an index table of linguistic texts, carrying out the analysis and stochastical indexing of the text-based documents and in forming a corresponding knowlegebase. A stochastically indexed user request is transformed into a multitude of new requests and the fragments of the text-based documents containing the word groups of the transformed requests are selected. Said fragments are used for forming a stochastically indexed semantic structure and the short response of the system based thereon. The relevance of the received short response to the request is checked by forming an interrogative sentence based thereon and by comparing said sentence with the request.

WO 03/100659 A1

[Продолжение на след. странице]

(57) Реферат: Изобретение может быть использовано при создании информационно-поисковых систем на базе Internet. При этом достигается возможность автоматического формирования знаний и извлечения их из текстовых документов, представленных на различных языках и в электронном виде, и интеллектуальная обработки текстовой информации и запросов пользователей. Для этого обеспечивают механизм самообучения системы правилам грамматического и семантического анализа в виде стохастически индексированной системы искусственного интеллекта. Формируют базы данных стохастически индексированных словарей и таблицы индексов лингвистических текстов. Производят анализ и стохастическое индексирование текстовых документов и формируют соответствующие базы знаний. Запрос пользователя преобразуют в стохастически индексированном виде во множество новых запросов и выбирают фрагменты текстовых документов со словосочетаниями преобразованного запроса, из которых формируют стохастически индексированную семантическую структуру, а на ее основе – краткий ответ системы. Релевантность полученного краткого ответа системы запросу проверяют путем формирования на его основе вопросительного предложения и сравнения его с запросом.

СПОСОБ СИНТЕЗА САМООБУЧАЮЩЕЙСЯ СИСТЕМЫ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ ТЕКСТОВЫХ ДОКУМЕНТОВ ДЛЯ ПОИСКОВЫХ СИСТЕМ

Область техники

Изобретение относится к области вычислительной техники, информационно-поисковых и интеллектуальных систем.

Изобретение предназначено для использования при создании информационно-поисковых и других информационных и интеллектуальных систем, работающих на базе Internet.

Предшествующий уровень техники

В настоящее время в системе Internet накоплен огромный объем информации по различным предметным областям и темам. В этой информации содержатся и постоянно обновляются всеобъемлющие сведения и знания. Однако доступ к ним со стороны многомиллионной пользовательской аудитории затруднен. Это обусловлено недостаточной эффективностью современных способов извлечения информации для поисковых систем. Известны способы извлечения информации для поисковых систем Yandex, Yahoo, Rambler. Известные способы обеспечивают выдачу текстовых документов по запросам пользователя из системы Internet.

Основными недостатками известных способов извлечения информации названных систем являются:

- сложность формализованных языков запросов;
- отсутствие аппарата семантического анализа содержания текстовых документов и их соответствия задаваемым вопросам;
- невозможность точного определения наличия в поисковом документе информации, указанной в запросе пользователя, а также выделения из объемных информационных источников конкретных сведений и знаний, необходимых пользователю.

В силу указанных недостатков при реализации информационно-поисковых процедур наряду с полезной передается много лишней, «шумовой» информации, которая плохо селектируется современными поисковыми системами. Это существенно повышает время поиска нужной информации, загружает каналы и серверы системы передачей и обработкой поискового шума.

Главная проблема состоит в том, что при этом и пользователь, задав запрос системе, получает большие объемы информации, часто не содержащей нужных сведений. Возникает необходимость ознакомиться с каждым полученным документом для определения наличия в нем требуемых данных. Это приводит к неоправданным временным и интеллектуальным затратам. Невозможность получения в реальном масштабе времени из огромных массивов Internet конкретных сведений и знаний, нужных пользователю для решения проблем различного характера, существенно снижает как ценность информации, так и эффективность работающих с ней поисковых систем.

Известен способ извлечения знаний и сведений по запросам пользователя из баз знаний, который реализован в интеллектуальной информационно-логической вычислительной системе, описанной в монографии: Насыпный В.В. Развитие теории построения открытых систем на основе информационной технологии искусственного интеллекта. М., 1994. - 248 с. (С.85-112). Указанный способ, основанный на стохастической информационной технологии, обеспечивает возможность эффективного поиска знаний и их обработки с использованием логического вывода в реальном масштабе времени. Это обусловлено тем, что в отличие от существующих способов обработки знаний, которые применяются в современных системах искусственного интеллекта, данный способ обеспечивает линейную зависимость времени поиска и логической обработки от объема знаний, необходимых для формирования ответа. Однако этот способ не дает возможности извлечения знаний из текстовых документов, что объясняется его ориентацией на обработку формализованной информации баз знаний, осуществляемой экспертами и инженерами по знаниям. Это делает невозможным использование данного способа для извлечения знаний из текстовых документов современных информационно-поисковых систем.

Известен также способ извлечения знаний из текстовых документов, описанный в работе: Насыпный В.В., Насыпная Г.А. Построение интеллектуальной информационно-поисковой системы. М.: Прометей, 2001. - 27 с. В основу способа положена стохастическая интеллектуальная информационная технология, которая обеспечивает проведение в реальном масштабе времени морфологического, синтаксического и семантического анализа больших объемов текстовой информации. Данная система может функционировать совместно с существующими информационно-поисковыми системами в качестве интеллектуальной надстройки над

ними, а также создавать поисковые системы нового поколения со своими стандартами стохастической индексации текстовых документов, протоколами информационного обмена и обработки запросов пользователя. Главными достоинствами указанного способа по сравнению со способами, реализованными в современных поисковых системах, являются:

- обработка запросов пользователя на естественном языке;
- поиск и выдача документов, достоверно содержащих полную информацию, релевантную запросу пользователя;
- выделение фрагментов текста в соответствии с запросом пользователя, содержащих сведения и знания по различным предметным областям, необходимым для решения конкретных проблем.

Основным недостатком данного способа является то, что наполнение баз знаний интеллектуальных систем, предназначенных для проведения морфологического, синтаксического, семантического анализа текста производится экспертами и требует длительных временных и технологических затрат. Поэтому создание подобных систем извлечения знаний из текстовых документов в интересах пользователей развитых стран, которые имеют национальные подсистемы в Internet с информацией на языке данной страны, требуют длительного времени. Вследствие этого указанный способ не может быть использован для создания на базе Internet многоязычных систем извлечения знаний из текстов. Это существенно затрудняет переход к индустрии знаний, которая бы основывалась на текстовой информации национальных поисковых систем и обеспечивала бы качественно новый информационный сервис в различных сферах – производственной, научной, образовательной, культурной и бытовой деятельности человека с учетом современных требований цивилизованного общества.

К другим недостаткам указанного способа можно отнести отсутствие возможности автоматического анализа новых слов, не входящих в состав словарей. В случае их появления в текстовых документах требуется участие экспертов при определении, к какой части речи относится новое слово, и его морфологических характеристик. Это делает невозможным автоматическое настраивание системы извлечения знаний на обработку текстовых документов по заданным новым темам. Отметим также, что для обеспечения эффективности извлечения знаний требуется комплексная обработка фрагментов текста из различных документов, основанная на анализе семантических связей с помощью логического вывода между указанными

фрагментами, а также на эквивалентных преобразованиях предложения данного текста. Эта функция также не реализована в рассматриваемом способе.

Раскрытие изобретения

Задачей изобретения является создание способа синтеза самообучающейся системы извлечения знаний из текстовых документов для поисковых систем для использования при создании глобальной индустрии знаний на базе Internet, не имеющего вышеуказанных недостатков. Достигаемым результатом является:

- возможность автоматического формирования знаний путем извлечения их из текстовых документов, представленных на различных языках в электронном виде для заполнения баз знаний;
- автоматический анализ новых слов и обновления словарей;
- эквивалентные преобразования запросов пользователей и предложений текстовых документов, обеспечивающие повышение эффективности извлечения знаний;
- самообучение указанных систем правилам грамматического и семантического анализа;
- интеллектуальная обработка текстовой информации и запросов пользователей с целью извлечения знаний на заданном иностранном языке.

Указанный технический результат достигается тем, что в способе синтеза самообучающейся системы извлечения знаний на заданном языке из текстовых документов поисковых систем

обеспечивают механизм самообучения в виде стохастически индексированной системы искусственного интеллекта, основанной на применении уникальных комбинаций двоичных сигналов стохастических индексов информации,

обеспечивают автоматическое обучение системы правилам грамматического и семантического анализа путем применения эквивалентных преобразований стохастически индексированных фрагментов текста, логического вывода и формирования из них связанных семантических структур и стохастического индексирования для представления в формате правил продукций,

производят морфологический анализ и стохастическое индексирование лингвистических текстов в электронном виде с одновременным автоматическим обучением системы правилам морфологического анализа,

производят морфологический и синтаксический анализ, а также стохастическое индексирование текстовых документов по заданной теме в электронном виде на

заданном языке с одновременным автоматическим обучением системы правилам синтаксического анализа,

производят семантический анализ стохастически индексированных текстовых документов по заданной теме в электронном виде с одновременным автоматическим обучением системы правилам семантического анализа,

формируют запрос пользователя на естественном заданном языке и представляют его в электроном виде после стохастического индексирования в форме вопросительного предложения,

преобразуют запрос пользователя в стохастически индексированном виде во множество новых запросов, эквивалентных исходному запросу,

в соответствии с запросом пользователя осуществляют предварительный выбор стохастически индексированных фрагментов текстовых документов в электронном виде, содержащих в совокупности все словосочетания преобразованного запроса,

формируют стохастически индексированную семантическую структуру с использованием указанных фрагментов текстовых документов,

на основе указанной структуры с помощью логического вывода, обеспечивающего связь стохастически индексированных элементов различных текстов, и эквивалентного преобразования текста формируют краткий ответ системы,

проверяют релевантность полученного краткого ответа системы запросу путем формирования на его основе вопросительного предложения, сравнения полученного вопросительного предложения с запросом,

при идентичности полученного вопросительного предложения и запроса принимают решение о релевантности краткого ответа системы запросу и представляют его на заданном языке.

Указанный технический результат достигается тем, что в способе синтеза самообучающейся системы извлечения знаний на любом из заданных иностранных языках из текстовых документов поисковых систем

обеспечивают механизм самообучения в виде стохастически индексированной системы искусственного интеллекта, основанной на применении уникальных комбинаций двоичных сигналов стохастических индексов информации для стохастической индексации и поиска фрагментов лингвистических текстов на заданном базовом языке, содержащих описание процедур грамматического и семантического анализа, и автоматического обучения системы правилам

грамматического и семантического анализа путем эквивалентных преобразований стохастически индексированных фрагментов текста, логического вывода и формирования из них связанных семантических структур, их стохастического индексирования для представления в формате правил продукций,

производят морфологический анализ и стохастическое индексирование лингвистических текстов на заданном базовом языке в электронном виде с одновременным автоматическим обучением системы правилам морфологического анализа, формированием базы данных стохастически индексированных словарей и формированием таблиц индексов лингвистических текстов для каждого из заданных иностранных языков, а также базы знаний морфологического анализа, содержащей полученные правила продукций для заданного базового языка и каждого из заданных иностранных языков,

производят морфологический и синтаксический анализ, а также стохастическое индексирование текстовых документов по заданной теме на каждом из заданных иностранных языков в электронном виде из поисковой системы с представлением их в виде таблиц индексов текстовых документов по заданной теме и записью в базы стохастически индексированных текстов с одновременным автоматическим обучением системы правилам синтаксического анализа с использованием стохастически индексированных лингвистических текстов на заданном базовом языке и формированием базы знаний синтаксического анализа для базового языка и каждого из заданных иностранных языков,

производят семантический анализ стохастически индексированных текстовых документов по заданной теме на заданном базовом языке в электронном виде с одновременным автоматическим обучением системы правилам семантического анализа и формированием базы знаний семантического анализа для базового языка и каждого из заданных иностранных языков,

формируют запрос пользователя на естественном заданном иностранном языке и представляют его в электроном виде после стохастического индексирования в форме вопросительного предложения, включающего вопросительное словосочетание и словосочетания, которые определяют семантику запроса,

преобразуют запрос пользователя в стохастически индексированном виде во множество новых запросов, эквивалентных исходному запросу на заданном иностранном языке,

в соответствии с запросом пользователя осуществляют предварительный выбор стохастически индексированных фрагментов текстовых документов на заданном иностранном языке в электронном виде, содержащих в совокупности все словосочетания преобразованного запроса,

формируют стохастически индексированную семантическую структуру на основе указанных фрагментов текстовых документов,

на основе сформированной стохастически индексированной семантической структуры с помощью логического вывода, обеспечивающего связь стохастически индексированных элементов различных текстов, и эквивалентного преобразования текста формируют краткий ответ системы, содержащий словосочетания в стохастически индексированном виде, которые определяют семантику запроса, а также группу слов ответа, соответствующую вопросительному словосочетанию запроса,

проверяют релевантность полученного краткого ответа системы запросу путем замены группы слов ответа на соответствующее вопросительное словосочетание в стохастически индексированном виде, получения стохастически индексированного вопросительного предложения, сравнения полученного вопросительного предложения с запросом и при идентичности полученного вопросительного предложения и запроса принимают решение о релевантности краткого ответа системы запросу и представляют его на заданном иностранном языке.

В случае неудачной попытки сформировать вопросительное предложение, идентичное запросу пользователя, запрашивают новые текстовые документы из поисковой системы для поиска ответа, релевантного запросу пользователя.

Дополнительно по запросу пользователя может быть сформирован полный ответ, содержащий более подробную информацию или совокупность конкретных знаний, при этом используют логический вывод для образования стохастически индексированной семантической структуры и необходимые эквивалентные преобразования указанной совокупности фрагментов текстов для получения стохастически индексированного нового текста, раскрывающего с возможной детализацией содержание полученного ранее краткого ответа.

При этом автоматическое обучение системы правилам морфологического анализа производят путем выделения в стохастически индексируемом тексте определенного набора словоформ каждого слова, получения стохастических индексов основы слова и заданного набора его окончаний или предлогов, произвольного доступа

по указанным индексам к стохастически индексированным лингвистическим текстам, выделения из них фрагментов, связывающих указанный набор окончаний слова или предлогов с соответствующей данному слову частью речи, а также с полным набором окончаний или предлогов, получаемых при склонении или спряжении, преобразования данных фрагментов в формат правил продукций путем их стохастического индексирования, обеспечивая при этом корректность каждого правила путем независимого его формирования на основе нескольких фрагментов из соответствующих лингвистических текстов, и получения таблицы индексов правил продукций для базы знаний морфологического анализа.

Кроме того, при стохастическом индексировании лингвистических текстов после определения части речи каждого слова с помощью правил базы знаний морфологического анализа заполняют базу данных стохастически индексированного словаря стохастическими индексами основы каждого очередного слова и полного набора его окончаний или предлогов, а при формировании таблиц индексов текстов осуществляют стохастическое преобразование информации и получение уникальных двоичных комбинаций индексов основ слов, их окончаний, предлогов, предложений, абзацев и названий текстов, которые помещают в таблицы индексов базы стохастически индексированных текстов с обеспечением связности между указанными индексами, определенной в исходном тексте и обеспечивающей его восстановление по таблице индекса.

Кроме того, автоматическое обучение системы правилам синтаксического анализа осуществляют путем поиска в стохастически индексированных лингвистических текстах фрагментов, описывающих порядок синтаксического разбора предложений, при этом реализуется логический вывод для получения стохастически индексированной семантической структуры, определяющей связь синтаксических элементов и структур с заданными частями речи слов, и формирования правил продукций, определяющих синтаксический разбор предложений по морфологическим характеристикам слов, обеспечивая при этом корректность каждого правила путем независимого его формирования на основе нескольких фрагментов из соответствующих лингвистических текстов, полученные правила заносят в базу знаний синтаксического анализа, по мере заполнения которой осуществляют ее стохастическое индексирование и представление в виде таблицы индексов.

Кроме того, автоматическое обучение системы правилам семантического анализа текста осуществляют путем формирования запроса к таблицам индексов лингвистических текстов по стохастическим индексам основ слов и частей речи, не точно определенных членов предложения, и получения ответа в виде фрагмента текста, описывающего семантические характеристики, которыми должны обладать слова для их соответствия данному конкретному члену предложения, и по полученному ответу, используя стохастический индекс основы данного слова и требуемые семантические характеристики, обращаются к таблицам индексов толковых словарей и энциклопедий общего и тематического назначения, при этом с помощью логического вывода делают попытку образовать стохастически индексированную семантическую структуру, связывающую данное слово и требуемые семантические характеристики, в положительном случае считают, что указанный член предложения определен точно, а фрагмент текста, релевантный запросу, преобразуют в правило продукций, обеспечивая при этом корректность каждого правила путем независимого его формирования на основе нескольких фрагментов из соответсвующих лингвистических текстов, которое включают в базу знаний семантического анализа, стохастически индексируют данную базу, представляют в виде таблицы индексов и применяют при семантическом анализе слов, как членов предложения, и отношений между словами, выраженных словосочетаниями.

После образования таблицы индексов каждого текста и завершения его морфологического, синтаксического и семантического анализа формируют стохастические индексы наименований частей речи, членов предложения и вопросов к ним, которые соответствуют каждому слову в составе предложений, и записывают указанные индексы в ячейки таблицы индексов данного текста, что позволяет при поиске фрагментов текста автоматически определять, к какой части речи, члену предложения относится каждое слово, и формировать вопросы к нему.

Затем, после получения всех таблиц индексов текстов, формируют таблицу индексов текстов по данной теме, строки которой поименованы неповторяющимися стохастическими индексами основ слов, а каждый столбец соответствует стохастическому индексу конкретного текста, при этом в ячейки таблицы записывают стохастические индексы абзацев, в которых в данном тексте содержится слово с соответствующим индексом основы, полученную таблицу индексов по данной теме

применяют для предварительного поиска фрагментов, содержащих определенную совокупность словосочетаний запроса.

При этом эквивалентные преобразования исходного запроса пользователя осуществляют с использованием синонимов, близких по смыслу слов, а также замены частей речи и членов предложения с сохранением смыслового содержания исходного запроса на основе применения стохастически индексированных правил морфологического, синтаксического и семантического анализа для получения эквивалентных структур словосочетаний вопросительного предложения запроса и сохранения семантической связи между ними.

Совокупность семантически связанных фрагментов текста, содержащих все слова запроса пользователя, формируют путем обращения по стохастическим индексам указанных основ слов к таблице индексов текстов по заданной теме, выбора стохастических индексов абзацев и соответствующих им текстов, содержащих в совокупности все словосочетания запроса, обращения по указанным индексам к таблице индексов каждого из выбранных текстов, логического вывода по таблицам индексов и эквивалентных преобразований текстов для образования стохастически индексированной семантической структуры, связывающей индексы группы слов ответа, соответствующего вопросительному словосочетанию запроса, а также все словосочетания запроса, определяющие семантику запроса и входящие в предварительно выбранные абзацы.

При этом успешно сформированная в процессе логического вывода стохастически индексированная семантическая структура, соответствующая запросу пользователя, принимается в качестве основы для формирования с использованием полученной совокупности фрагментов текста вопросительного предложения, идентичного запросу пользователя, которое образуют путем эквивалентного преобразования стохастических индексов основ слов запроса и их окончаний с помощью правил баз знаний для обеспечения требуемых семантических характеристик каждого словосочетания текстового фрагмента, входящего в состав запроса, а также с использованием логического вывода на транзитивных зависимостях между словосочетниями для объединения их в единое вопросительное предложение, идентичное запросу пользователя, которое содержит группу слов ответа, соответствующую вопросительному словосочетанию запроса.

Корректность краткого ответа может быть обеспечена путем формирования нескольких идентичных стохастически индексированных семантических структур упомянутого ответа на основе различных, предварительно выбранных стохастически индексированных фрагментов текстовых документов.

В процессе поиска и формирования ответа с использованием таблиц индексов текстовых документов самообучение системы осуществляют путем формирования индексированных текстовых элементов, связывающих запрос и релевантный краткий ответ, для получения базы знаний, содержащей элементы типа «запрос – ответ», которую стохастически индексируют, представляют в виде таблицы индексов и применяют при грамматическом и семантическом анализе предложений текста, а также при формировании ответов на повторяющиеся запросы пользователей, содержащиеся в указанной индексированной базе знаний.

При этом для формирования полного ответа, содержащего знания, релевантные запросу пользователя, на основе краткого ответа с помощью логического вывода по таблицам индексов, использованных при получении фрагмента текста, формируют стохастически индексированную семантическую структуру, связывающую группу слов ответа со стохастическими индексами основ слов предложений, поддерживающих транзитивную зависимость, обеспечивающую в своей совокупности полное раскрытие содержания краткого ответа в рамках сформированного фрагмента текста, затем с помощью эквивалентных преобразований предложений на основе указанной стохастически индексированной семантической структуры получают единый связанный текст полного ответа.

Эквивалентное преобразование стохастически индексированных фрагментов текста производят путем представления каждого предложения в виде совокупности стохастически индексированных словосочетаний, которые преобразуют с использованием правил баз знаний морфологического, синтаксического и семантического анализа путем эквивалентного преобразования стохастических индексов основ однокоренных слов, их окончаний и предлогов для образования новых частей речи или членов предложения с обеспечением неизменности связи указанных словосочетаний в рамках стохастически индексированной семантической структуры каждого предложения и согласования указанных предложений между собой при образовании из них нового фрагмента текста.

При появлении в процессе стохастического индексирования текстовых документов в индексируемом тексте нового слова, не содержащегося в словаре стохастически индексированных слов и в лингвистических текстах, находят в данном словаре однокоренное слово с указанным новым словом, а в базе знаний морфологического анализа находят правила для эквивалентного преобразования найденного в словаре однокоренного слова в новое слово, при этом по виду эквивалентного преобразования определяют часть речи, к которой относится новое слово и все его словоформы, получаемые при склонении или спряжении, а при отсутствии однокоренных слов в словаре выбирают из текста определенный набор словоформ нового слова, по предлогам или окончаниям которых с помощью стохастически индексированного словаря или правил продукций морфологического анализа определяют часть речи, к которой оно относится, и полный набор его словоформ, получаемых при склонении или спряжении.

При этом для одновременного извлечения знаний из текстовых документов на заданных иностранных языках сначала осуществляют автоматическое обучение системы правилам морфологического, синтаксического, семантического анализа для заданного базового языка, производят формирование базы стохастически индексированного словаря и баз знаний морфологического, синтаксического, семантического анализа с использованием стохастически индексированных лингвистических текстов на заданном базовом языке, с помощью сформированных баз осуществляют автоматическое формирование запросов для автоматического обучения системы любому из заданных иностранных языков, при этом производят предварительный выбор по автоматически сформированным запросам фрагментов лингвистических текстов на базовом языке, содержащих знания, необходимые для изучения заданного иностранного языка, эквивалентные преобразования указанных текстов, формирование стохастически индексируемых семантических структур и логический вывод на заданных структурах для формирования ответов, релевантных автоматическим запросам, которые используют для формирования баз знаний морфологического, синтаксического и семантического анализа для любого из заданных иностранных языков, обеспечивающих извлечение знаний из текстовых документов на заданном иностранном языке.

Краткое описание чертежей

Изобретение поясняется на примере, иллюстрируемом Фиг.1, где показана структурная схема интеллектуальной самообучающейся системы извлечения знаний из текстовых документов для поисковых систем; а также следующими таблицами:

Таблица 1- Фрейм предложения,

Таблица 2 - Индексы текста,

Таблица 3 - Индексы текстов по данной теме.

Предпочтительный вариант осуществления изобретения

Ниже приведены определения терминов, используемых в настоящем описании.

База знаний – один или несколько специальным образом организованных файлов, хранящих систематизированную совокупность понятий, правил и фактов, относящихся к некоторой предметной области.

Вопросительное словосочетание – словосочетание с вопросительным местоимением или наречием в роли вопросительного слова, связанного с главным словом словосочетания (именем или глаголом).

Грамматический анализ – анализ морфологический и синтаксический.

Знания – новая текстовая информация, не содержащаяся в явном виде в текстовых документах, которая автоматически формируется системой с использованием эквивалентных преобразований и логического вывода в виде ответа, релевантная запросу пользователя и направленная на решение его задач в соответствии с запросом.

Лингвистические тексты – учебно-методические, научные, справочные (толковые словари, энциклопедии) и другие тексты, предназначенные для изучения данного языка.

Логический вывод – метод обработки знаний, имитирующий процесс рассуждений человека, который на основе отдельных языковых единиц позволяет синтезировать семантическую структуру с определенным смысловым содержанием.

Морфологический анализ – это разбор слов предложения для определения морфологического состава с последующим уточнением характеристик отдельных слов, относящихся к той или иной части речи, при этом вначале указываются постоянные морфологические признаки слова, не зависящие от его позиции в предложении, затем анализируется грамматическая форма слова, связанная с его склонением или спряжением.

Основа слова - часть слова, выражающая его лексическое значение, при этом в склоняемых и спрягаемых словах имеются основа и окончание, а остальные слова содержат только основу.

Поисковая система – система, выполняющая автоматический поиск информации по ключевым словам, темам и т.д.

Правила продукций – форма представления знаний в виде сложноподчиненного предложения «Если (условие), то (заключение)», в котором условие содержит различные словосочетания, включающие предикативные и другие виды отношений между объектами предметной области, объединенные логическими связками «и», а заключение содержит словосочетание или совокупность словосочетаний, определяющих семантическое следствие, которое истинно, или действие, которое активизируется, если истинны все словосочетания условия.

Релевантность – мера, определяющая, насколько полно тот или иной документ отвечает критериям, указанным в запросе пользователя.

Семантическая структура – форма связи отдельных языковых единиц различных предложений с учетом видов отношений между ними, выражающая определенное смысловое содержание анализируемого текста.

Семантический анализ - анализ смысла, значения отдельных языковых единиц: слов, словосочетаний предложения, их соотнесенности с определенными видами отношений между объектами предметной области и явлениями действительности.

Синтаксический анализ - это разбор слов предложения для определения синтаксического состава с последующим уточнением характеристик отдельных слов, словосочетаний, их типов, форм связи между словами в словосочетании и предложении, строения предложений, структурных типов предложений.

Система искусственного интеллекта – программно-техническая система, содержащая в качестве основы подсистему логического вывода, базы знаний, а также в зависимости от класса другие программно-аппаратные средства искусственного интеллекта и предназначенная для поддержки интеллектуальной деятельности человека или его замены в ряде процессов управления.

Склонение слова – изменение существительных по падежам (для большинства имен и по числам), а для прилагательных и других согласуемых слов также по родам.

Словосочетание - это синтаксическая единица, образующаяся соединением двух или более слов на основе подчинительной связи – согласования, управления или

примыкания – и тех лексико-грамматических отношений, которые порождаются этой связью.

Словоформа – данное слово в данной грамматической форме.

Спряжение слова – изменение глагола по лицам, числам, временам и наклонениям, а в прошедшем времени и в сослагательном наклонении в единственном числе также по родам.

Эквивалентное преобразование – замена отдельных языковых единиц на другие с обеспечением их связи в рамках семантической структуры предложения или в определенной совокупности предложений текста, способных выражать то же смысловое содержание.

Рассмотрим более подробно реализацию предложенного способа на примере построения и функционирования интеллектуальной самообучающейся системы извлечения знаний для поисковых систем (ИССИЗ), представленной на Фиг.1. Упомянутая стохастически индексированная система искусственного интеллекта включает:

- многоязычный лингвистический процессор (1);
- подсистему стохастического индексирования текстовых документов и выделения фрагментов текстов (2);
- подсистему управления режимом самообучения и извлечения знаний (3);
- интерпретатор стохастически индексированных текстов и правил продукций (4);
- подсистему эквивалентных преобразований текста (5);
- подсистему логического вывода (6);
- базу данных стохастически индексированных словарей базового и новых слов (7);
- базу стохастически индексированных лингвистических текстов (8);
- базу знаний «запрос-ответ» (9);
- базу стохастически индексированных текстовых документов по заданным темам (10);
- базу стохастически индексированных словарей иностранных слов (11);
- базу знаний морфологического анализа (12);
- базу знаний синтаксического анализа (13);
- базу знаний семантического анализа (14);
- базу стохастически индексированных словосочетаний (15).

Указанная система основана на использовании стохастического преобразования и индексирования символьной информации, формирования таблиц индексов правил

продукций для управления режимом самообучения и индексов текстов. Она обеспечивает доступ по стохастическим индексам к фрагментам текстовой информации, логический вывод и эквивалентные преобразования текста с использованием стохастически индексированных правил для извлечения знаний из выделенных фрагментов текста и представления их в формате правил продукций или в виде ответов на запросы пользователей.

Создание ИССИЗ предполагает разработку механизма самообучения системы правилам морфологического, синтаксического и семантического анализа текстовой информации на основе лингвистических текстов. Указанные тексты содержат словари общеупотребительных слов, тематические словари, словари синонимов, толковые словари, учебно-методические тексты по грамматике заданных языков и др.

Общение пользователя с системой осуществляется через многоязычный лингвистический процессор (1). Он обеспечивает ввод запросов на естественном языке и выдачу ответов, формируемых системой. При этом обмен информации между пользователем и системой может осуществляться на заданных языках. Кроме этого лингвистический процессор (1) по команде подсистемы (3) управления режимом самообучения и извлечения знаний обеспечивает взаимодействие с подключенной к ИССИЗ поисковой системой. Цель этого взаимодействия - ввод по запросу подсистемы (3) новых текстовых документов из поисковой системы на заданном языке по определенной теме для их последующей обработки. Многоязычный лингвистический процессор (1) также обеспечивает ввод в систему лингвистических текстов на заданном языке в электронном виде.

Морфологический анализ лингвистических текстов и автоматическое обучение системы правилам морфологического анализа производят по команде подсистемы (3) управления режимом самообучения и извлечения знаний в процессе формирования базового словаря и записи его в базу данных (7) стохастически индексированных словарей базового и новых слов. Эти функции проводят одновременно с индексированием лингвистических текстов с помощью подсистемы (2) стохастического индексирования текстовых документов и выделения фрагментов текстов.

Для формирования стохастически индексированного базового словаря используют словарь общеупотребительных слов в электронном виде, который вводят в систему через многоязычный лингвистический процессор (1) и определяют по заданным словоформам этого словаря часть речи каждого слова, его основу, и

соответствующие наборы окончаний. Основу данного слова стохастически индексируют с помощью подсистемы (2) стохастического индексирования текстовых документов и выделения фрагментов текстов и записывают в базу данных (7) стохастически индексированных словарей базового и новых слов в таблицу стохастически индексированного базового словаря в столбец индексов основ слов.

В результате описанной обработки слов указанного словаря в многоязычном лингвистическом процессоре (1) получают стохастические индексы основ всех слов и сами основы, а также определенный набор окончаний, которые заносят в базу данных (7) стохастически индексированных словарей базового и новых слов.

Стохастически индексированный базовый словарь, записанный в базу данных (7) стохастически индексированных словарей базового и новых слов, имеет несколько форматов таблиц, каждая из которых соответствует определенной части речи. В заголовке таблиц содержатся графы, включающие наименования морфологических характеристик (род, число, падеж, лицо, время и т.д.), а также вопросы, которые соответствуют словоформам данного слова, получаемым при его склонении или спряжении. При этом каждой основе соответствует строка, содержащая окончания указанных словоформ данного слова. Отметим, что в начале заполнения стохастически индексированного базового словаря, известно только несколько словоформ каждого слова, а именно те, которые приведены в словаре общеупотребительных слов. Нахождение остальных словоформ и соответствующих им окончаний для заполнения таблиц стохастически индексированного базового словаря производят в режиме автоматического обучения системы правилам морфологического анализа после первоначального индексирования соответствующих лингвистических текстов.

В основу этого механизма положено введение нового способа стохастической индексации текстовых документов, который реализуется в подсистеме (2) стохастического индексирования текстовых документов и выделения фрагментов текстов. Процедура базируется на функциях стохастического преобразования символьной информации и формирования стохастических индексов в виде уникальных двоичных комбинаций основ слов, предложений, абзацев и названий текстовых документов, включая библиографические данные. При этом одновременно со стохастическим преобразованием символьной информации, формированием стохастических индексов $\{I_{\xi_i}^{(u)}\}$ основ слов, предложений $\{I_{\xi_i}^{(p)}\}$, абзацев $\{I_{\xi_i}^{(a)}\}$ и

названия текста $I_{\xi j}^{(t)}$, который находится в обработке, производится заполнение фреймов каждого предложения (Таблица 1) и формирование таблицы индексов данного текста (Таблица 2).

Указанный фрейм (Таблица 1), который формируется в подсистеме (2) стохастического индексирования текстовых документов и выделения фрагментов текстов, содержит десять уровней (строк) слотов (ячеек). Эти уровни слотов заполняются в процессе стохастического индексирования текста, а также при выполнении морфологического, синтаксического и семантического анализа каждого предложения.

При стохастическом индексировании лингвистических текстов в слоты первого уровня записываются стохастические индексы основ слов $\{I_{\xi i j}^{(u)}\}$ и их окончания. Слоты второго уровня содержат слова в порядке следования в данном предложении с номером **i**. При этом предлоги, частицы, союзы и знаки препинания заносят в слоты тех слов, с которыми они связаны. Для заполнения слотов третьего уровня используются стохастические индексы основ слов $\{I_{\xi ij}^{(u)}\}$ и их окончания, записанные в слоты первого уровня.

По индексам основ слов производят доступ к строкам соответствующих таблиц стохастически индексированного базового словаря, поименованным идентичными индексами для определения части речи, к которой относится данное слово. Указанную информацию из базы данных (7) стохастически индексированных словарей базового и новых слов записывают в слоты третьего уровня фрейма предложения, соответствующие словам слотов второго уровня.

Запись в слоты третьего уровня фрейма характеристик частей речи, а также заполнение слотов уровней с четвертого по десятый производят в процессе дальнейшего морфологического и синтаксического анализа текста, который осуществляют одновременно с обучением системы правилам морфологического и синтаксического анализа. Этот процесс будет рассмотрен ниже.

На основе получения фреймов предложений текста с заполненными первыми четырьмя уровнями слотов в подсистеме (2) стохастического индексирования текстовых документов и выделения фрагментов текстов осуществляют формирование таблицы индексов данного текста.

Таблица 2 индексов текста представляет собой таблицу, строки которой поименованы стохастическими индексами $\{I_{\xi i}^{(u)}\}$ основ слов, столбцы обозначены индексами абзацев $\{I_{\xi j}^{(a)}\}$ в порядке их появления в тексте, а ячейки, расположенные на пересечении соответствующих столбцов и строк, содержат индексы списков $\{I_{\xi i j}^{(s)}\}$. При этом сама информация, которая содержится в каждом списке, поименованная $\{I_{\xi i j}^{(s)}\}$, записана в отдельном файле и в общем случае должна включать следующие данные:

$\{I_{\xi i}^{(p)}\}$ – индекс предложения, в которое входит данное слово;

$N_i^{(n)}$ – номер предложения, в которое входит данное слово;

$(u_i u_j)$ - окончание, которое имеет данное слово в предложении $(I_{\xi i}^{(p)} N_i^{(n)})$;

$I_{\xi j-1}^{(u)}$ - индекс предшествующего слова в предложении или абзаце текста, при этом, если $I_{\xi j}^{(u)}$ - первое слово в предложении (абзаце), то после индекса $I_{\xi j-1}^{(u)}$ ставится точка. $I_{\xi j-1}^{(u)}$ может соответствовать слову, завершающему предыдущее предложение в рамках данного абзаца или предыдущего абзаца. Если после $I_{\xi j-1}^{(u)}$ ставится запятая, то это означает, что $I_{\xi j}^{(u)}$ может начинать причастный или деепричастный оборот, придаточное предложение или простое предложение в составе сложного;

$I_{\xi(j+1)}^{(u)}$ - индекс последующего слова в предложении, абзаце, тексте, при этом, если $I_{\xi j}^{(u)}$ - завершающее слово в предложении (абзаце), то перед $I_{\xi j-1}^{(u)}$ ставится точка. $I_{\xi j-1}^{(u)}$ может соответствовать слову, начинающему новое предложение данного абзаца или последующего абзаца. Если перед $I_{\xi j-1}^{(u)}$ ставится запятая, то это означает, что $I_{\xi j}^{(u)}$ может завершать деепричастный, причастный обороты или простое предложение в составе сложного;

$I_{\xi j}^{(vu)}$ - индекс вопроса к данному слову, как к члену предложения;

$I_{\xi j}^{(pu)}$ - индекс наименования члена предложения, которому соответствует данное слово;

$I_{\xi j}^{(vpru)}$ - индекс вопроса, которому соответствуют деепричастный, причастный обороты или придаточное предложение, которое начинает $I_{\xi j}^{(u)}$;

$I_{\xi j}^{(pru)}$ - индекс наименования члена предложения, которому соответствуют причастный, деепричастный обороты или придаточное предложение, начинающее $I_{\xi j}^{(u)}$.

Указанные индексы и символы соответствуют слову с основой $I_{\xi i}^{(u)}$ в составе одного из предложений $I_{\xi j}^{(p)}$ абзаца $I_{\xi j}^{(a)}$ и имеют заданный формат, определяющий расположение индексов и символов в составе данной группы. Если отдельные индексы отсутствуют, то вместо них на соответствующей позиции ставится знак «пробел». Если данное слово $I_{\xi i}^{(u)}$ входит в **n** предложений $\{I_{\xi i}^{(p)}\}$ абзаца $I_{\xi i}^{(a)}$, то указанных групп в составе списка также будет **n**.

Отметим, что первые шесть индексов списка $I_{\xi i\, j}^{(s)}$ формируются в ходе стохастического индексирования текста. При этом по индексу $I_{\xi i}^{(u)}$ основы путем обращения к стохастически индексированному базовому словарю всегда можно определить, к какой части речи относится указанное слово. Остальные данные списка $I_{\xi i\, j}^{(s)}$ определяются после заполнения уровней четыре- десять фреймов предложений текста в процессе дальнейшего морфологического и синтаксического разбора, которые реализуются одновременно с самообучением системы правилам грамматического анализа предложений.

После стохастического индексирования всех лингвистических текстов, включая тексты, содержащие описания грамматического разбора предложений, их записывают в базу (8) стохастически индексированных лингвистических текстов и переходят к

формированию правил морфологического анализа текста одновременно с заполнением базы данных (7) стохастически индексированных словарей базового и новых слов.

С этой целью из каждой таблицы стохастически индексированного базового словаря, которая содержит основы слов, относящихся к данной части речи, выбирают стохастический индекс основы каждого слова и заданного набора его окончаний или предлогов. Затем осуществляют произвольный доступ по указанным индексам к базе (8) стохастически индексированных лингвистических текстов для выделения из них фрагментов, связывающих индекс части речи и указанный набор окончаний слова или предлогов с соответствующими данной части речи полным набором окончаний, предлогов или вопросов, получаемых при склонении или спряжении. После этого данный фрагмент текста поступает в интерпретатор (4) стохастически индексированных текстов и правил продукций, в котором формируют стохастически индексированную семантическую структуру в виде совокупности словосочетаний каждого предложения, входящего в данный фрагмент:

$$S: \{(I_{\xi i}^{(u)} I_{\xi i}^{(r)} I_{\xi i}^{(z)}) \longrightarrow (I_{\xi j}^{(u)} I_{\xi j}^{(r)} I_{\xi j}^{(z)})\}, \qquad (1)$$

где $I_{\xi i}^{(u)} I_{\xi j}^{(u)}$ - стохастические индексы соответственно главного и зависимого основ слов данного словосочетания, $I_{\xi i}^{(r)} I_{\xi j}^{(r)}$ - стохастические индексы частей речи главного и зависимого слов указанного словосочетания, $I_{\xi i}^{(z)} I_{\xi j}^{(z)}$ - стохастические индексы соответственно морфологических характеристик частей речи главного и зависимого слов данного словосочетания, а знак $\longrightarrow$ определяет связь между главным и зависимым словами данного словосочетания.

Основным связующим звеном каждой стохастически индексированной семантической структуры, представленной выражением (1), является глагол, который определяет семантику связей внутри данной структурной схемы. Связь между различными стохастически индексированными семантическими структурами (1), входящими в разные предложения, осуществляется при наличии в них идентичных словосочетаний, их синонимов, повторения главных слов или применения во втором предложении местоимения, соответствующего одному из словосочетаний первого предложения, а также местоимения в сочетании с главным словом. В соответствии с этим находят предложения или части предложений, в которых стохастически

индексированная семантическая структура, содержащая индексированные исходные данные запроса, соответствующим образом связана со стохастически индексированной семантической структурой с индексированными данными ответа. При этом для определения семантики глаголов производят обращение по стохастическим индексам их основ к базе (8) стохастически индексированных лингвистических тексов для доступа к таблицам индексов словарей синонимов.

Если первая и вторая структурные схемы связаны между собой словосочетанием, содержащим определяемую часть речи, а значения глаголов, связанные с данной частью речи, идентичны или синонимичны глаголам запроса и предполагаемого ответа, то указанные структурные схемы поступают в подсистему (5) эквивалентных преобразований текста. В подсистеме (5) производится преобразование двух указанных семантических структурных схем в единую стохастически индексированную семантическую структуру правила продукций, которая содержит условие, включающее запрос, и заключение (ответ). Указанная стохастически индексированная семантическая структура имеет в общем случае следующий вид:

$$P: I_{\xi 1}^{(su)} \wedge I_{\xi 2}^{(su)} \wedge I_{\xi 3}^{(su)} \wedge \ldots \wedge I_{\xi m}^{(su)} \Longrightarrow$$

$$I_{\xi 1}^{(su)} \wedge I_{\xi 2}^{(su)} \wedge_{\xi 3}^{(su)} \wedge \ldots \wedge I_{\xi n}^{(su)} \qquad (2)$$

где $I_{\xi i}^{(su)}$ - является стохастическим индексом соответствующего словосочетания

$I_{\xi i}^{(su)}: (I_{\xi i}^{(u)} I_{\xi i}^{(r)} I_{\xi i}^{(z)}) \longrightarrow (I_{\xi j}^{(u)} I_{\xi j}^{(r)} I_{\xi j}^{(z)})$ из выражения (1), а секвенция $\Longrightarrow$ истолковывается в обычном логическом смысле как знак логического следования заключения, находящегося в правой части выражения (2), из условия в левой части выражения (2), если все словосочетания условия являются истинными (соответствуют исходным данным запроса). Отметим, что корректность каждого правила обеспечивается при этом путем независимого формирования описанным выше порядком идентичных стохастически индексированных семантических структур (2) на основе нескольких фрагментов из соответствующих лингвистических текстов.

Каждое правило продукций, сформированное в подсистеме (5) эквивалентных преобразований в виде выражения (2), поступает в интерпретатор (4) стохастически

индексированного текста и правил продукций, где осуществляют преобразование данного выражения (2) в текстовый формат правил продукций, представленный в виде «Если (условие), то (заключение)». Полученное правило в индексированном виде поступает в базу знаний (11) морфологического анализа. Порядок синтеза баз знаний, содержащих стохастически индексированные правила, будет описан ниже.

При формировании правил морфологического анализа текста одновременно с заполнением базы данных (7) стохастически индексированных словарей базового и новых слов первая стохастически индексированная семантическая структура (1) (структурная схема) содержит стохастические индексы основы слова, обозначающие часть речи и заданный набор его окончаний или предлогов. Вторая структурная схема (1) связывается с первой через идентичный индекс части речи и определяет полный набор окончаний, предлогов, вопросов, получаемых при склонении или спряжении данной части речи.

Путем обращения описанным выше порядком к таблицам индексов словарей синонимов, соответствующих лингвистическим текстам базы (8) стохастически индексированных лингвистических текстов, определяют соответствие семантики глаголов первой и второй семантических структур запросу и предполагаемому ответу. Затем определяют словосочетание, связывающее первую и вторую стуктуры. При положительном результате две части указанного фрагмента текста поступают в подсистему (5) эквивалентых преобразований текста, затем в интерпретатор (4) стохастически индексированных текстов и правил продукций. В результате осуществляют преобразование данного фрагмента в формат правил продукций, представленный в виде «Если (условие), то (заключение)». При этом в условие правила входят индексы словосочетаний, связывающих часть речи и заданный набор окончаний слова или предлогов, расположенных в формате словаря и определяющих изменения словоформы при склонении или спряжении данного слова. Заключение содержит полный набор окончаний, предлогов и вопросов, получаемых при склонении или спряжении данного слова как соответствующей части речи. Сформированное правило продукций записывают в базу знаний (11) морфологического анализа. После завершения формирования правил, определяющих части речи, по команде подсистемы (3) управления режимом самообучения и извлечения знаний переходят к синтезу правил эквивалентных преобразований однокоренных слов. Здесь используется предварительно записанное в базу знаний (11) морфологического анализа общее

правило преобразования частей речи, основанное на применении таблиц стохастически индексированного базового словаря и выборе соответствующих фрагментов лингвистических текстов, которые описывают порядок образования одной части речи на базе другой однокоренной части речи:

«Если требуется преобразовать одну часть речи в другую,
то сначала выделяем основу первой части речи,
обращаемся к формату стохастически индексированного базового словаря,
ищем вторую часть речи, основа которой имеет общую часть, включающую корень (возможно два, возможно с приставкой, возможно с чередованием, добавлением, исключением отдельных гласных или согласных), с основой первой части речи,
после выделения корня, используя основу этих частей речи, выделяем их суффиксы,
затем, путем обращения по стохастическим индексам основ слов частей речи к таблицам индексов лингвистических текстов выбираем фрагмент, в котором описан соответствующий способ преобразования одной части речи в другую, и проверяем по формату словаря, каким способом образована основа второй части речи по отношению к основе первой (заменой, отбрасыванием, прибавлением суффиксов),
далее определяем, соответствует ли данный способ замены части речи требуемому способу образования второй части речи из первой части речи,
в положительном случае принимаем вторую часть речи в качестве вновь образованной».

В процессе преобразования конкретных слов с использованием общего правила на его основе формируется соответствующее частное правило с указанием преобразуемых частей речи, суффиксов и способа образования одной части речи из другой. Это происходит в интерпретаторе (4) стохастически индексированных текстов и правил продукций и в подсистеме (5) эквивалентных преобразований текста. Описанным выше порядком осуществляют преобразование данного фрагмента сначала в единую стохастически индексированную семантическую структуру правила продукций (2), а затем в формат правил продукций, представленный в виде «Если (условие), то (заключение)». Эти правила после стохастического индексирования заносятся в базу знаний (11) морфологического анализа.

Если при индексировании очередного текстового документа появляется новое слово, основа которого не содержится в базовом словаре, то переходят к процедуре определения части речи нового слова и его окончаний при склонении или спряжении.

Для начала процесса определения, к какой части речи относится новое слово, выделяют из текста не менее двух различных словоформ этого слова, путем их сравнения определяют неизменяемую часть, которая предположительно является основой нового слова, и его окончание. После этого определяют, есть ли в формате базового словаря слова, имеющие общий корень (возможно с приставкой) с новым словом. Корнем является общая, нечленимая часть основ родственных слов (содержащая не менее двух букв, включая одну гласную), которую при добавлении приставок, суффиксов и окончаний используют для образования однокоренных частей речи. В соответствии с этим выделение общего корня производят путем сравнения основы нового слова и основ слов из формата базового словаря до тех пор, пока не найдут общую неделимую часть двух сравниваемых основ - нового слова и очередного слова из базового словаря.

После этого производят обращение к базе знаний (12) морфологического анализа для выбора правила, позволяющего определить, к какой части речи относится новое слово. С этой целью используют соответствующее правило эквивалентных преобразований.

Чтобы использовать правила эквивалентных преобразований для определения части речи нового слова, полагают, что вторая часть речи в общем правиле эквивалентных преобразований, приведенном выше, относится к новому слову и является неизвестной, при этом первая часть речи, имеющая с ним общий корень, найдена в базовом словаре и поэтому известна. Затем проверяют, возможно ли с помощью преобразований, описанных в правиле, получить из основы известной части речи основу нового слова, часть речи которого неизвестна. При этом используется семейство конкретных правил, полученных на основе общего правила и содержащихся в базе знаний (12) морфологического анализа, которые позволяют преобразовать известную (первую) часть речи в другие части речи. Если в результате использования одного из правил удастся получить основу нового слова, то часть речи, к которой оно относится, станет известной – оно будет соответствовать второй части речи, указанной в правиле. При этом с использованием правил продукций базы знаний (12) морфологического анализа можно более подробно определить характеристики каждой части речи. Например, если при морфологическом анализе текстов на русском языке правила базы знаний (12) морфологического анализа позволяют определить не только часть речи нового слова, но и окончание имени (сущ., прил.) в им.п., ед.ч., то,

следовательно, они дают возможность уточнить, к какому типу склонения (1, 2, 3) относится новое слово. Для имен существительных, прилагательных, порядковых числительных, некоторых видов местоимений, а также причастий это позволяет точно определить полный набор их окончаний, получаемых при склонении. В данном случае для указанных частей речи достаточно найти в формате словаря соответствующее им слово, имеющее в им.п. ед.ч. такое же окончание, как в новом слове. Полный набор окончаний указанных частей речи будет соответствовать набору окончаний нового слова, которые записывают в формат словаря новых слов вместе с его основой. После этого формируют стохастический индекс основы, а все полученные характеристики нового слова записывают в формат словаря новых слов.

Если новое слово является глаголом, то после выделения его основы описанным выше порядком и обращения к базе знаний (12) морфологического анализа с помощью соответствующего правила определяют его часть речи и находят инфинитив. По суффиксу данного инфинитива ***(-ть*** или ***–ти),*** обращаясь к формату базового словаря, находят глагол, который имеет в неопределенной форме такой же суффикс ***(-ть*** или ***–ти).*** При этом полный набор окончаний данного глагола, полученных после его спряжения и записанных в формате словаря, предположительно выбирают в качестве полного набора окончаний нового глагола. Для более точного определения, к какому типу спряжения (1, 2) относится данный глагол и, соответственно, для уточнения полного набора его окончаний в процессе индексирования текста находят предложение, в котором данный глагол представлен в форме 3-его л. мн.ч. Для этого находят предложение, в котором есть подлежащее, выраженное существительным (местоимением) во мн.ч., которое координирует со сказуемым, выраженным данным глаголом с личным окончанием ***–ут /–ют*** (1 спряжение) или ***–ат /–ят*** (2 спряжение). По личному окончанию отмеченного глагола в формате базового словаря находят глагол, имеющий идентичное с ним окончание в 3-ем л. мн.ч. При этом полный набор окончаний данного глагола принимают в качестве полного набора окончаний нового глагола и записывают вместе с его основой в формат словаря новых слов. После получения стохастического индекса основы нового глагола всю указанную информацию записывают в формат словаря новых слов.

В процессе индексирования текста при появлении различных словоформ новых слов, не содержащихся в базе данных (7) стохастически индексированных словарей базового и новых слов, путем сравнения указанных словоформ в подсистеме (2)

стохастического индексирования текстовых документов и выделения фрагментов текстов осуществляют выделение основы нового слова и определенного набора его окончаний. Затем формируют стохастический индекс основы нового слова и вместе с его окончаниями заносят в формат словаря новых слов базы данных (7) стохастически индексированных словарей базового и новых слов. После обработки заданного набора словоформ данного слова и соответственно заполнения формата словаря с различными видами его окончаний производят обращение к таблице индексированного базового словаря. Данный словарь после заполнения содержит индексы и основы общеупотребительных слов, а также все виды окончаний различных частей речи и их типов, относящихся к данному слову, которые получены при его склонении или спряжении с указанием характеристик частей речи. Запрос к словарю содержит стохастический индекс основы данного слова, саму основу, а также все виды окончаний, которые имели словоформы этого слова при обработке текстовых документов. В базе данных (7) стохастически индексированных словарей базового и новых слов по окончаниям данного слова, используя формат словаря, находится слово, имеющее такие же окончания среди полного набора окончаний. Это означает, что новое слово относится к такой же части речи, как и слово в словаре, имеющее идентичные окончания. После определения части речи, к которой относится новое слово, всю информацию, входящую в запрос, заносят в словарь новых слов в установленном формате. Одновременно с этим в интерпретаторе (4) стохастически индексированных текстов и правил продукций и в подсистеме (5) эквивалентных преобразований текста описанным выше порядком осуществляют преобразование данного фрагмента сначала в единую стохастически индексированную семантическую структуру (2) правила продукций, а затем в формат правил продукций, представленный в виде «Если (условие), то (заключение)».

В результате формируется правило продукций, в условие которого входит заданный набор окончаний данного слова, а заключение содержит наименование части речи данного слова, имеющего приведенные в условии окончания, а также расположенный в формате словаря полный набор окончаний, которые определяют изменения словоформы при склонении или спряжении данного слова. Кроме этого в заключение содержатся вопросы к словоформам данной части речи при ее склонении или спряжении, которые расположены в порядке, определяемом форматом словаря.

Таким образом, в процессе обработки текстов, содержащих новые слова, которые представлены в своих различных словоформах, производится автоматическое определение их части речи, заполнение формата словаря новых слов в базе данных (7) стохастически индексированных словарей базового и новых слов, а также обучение системы правилам морфологического анализа. Эти правила заносятся в базу знаний (12) морфологического анализа.. По мере заполнения базы знаний (12) и ее стохастического индексирования описанным ниже порядком она наряду с форматом стохастически индексированного базового словаря используется для определения, к какой части речи относится новое слово и его характеристики, если оно не содержится в формате словаря новых слов..

После завершения морфологического анализа и стохастического индексирования лингвистических текстов, формирования базы знаний (12) морфологического анализа, базы (8) стохастически индексированных лингвистических текстов, а также базы (7) стохастически индексированных словарей базового и новых слов переходят к стохастическому индексированию текстов по заданной теме с одновременным автоматическим обучением системы правилам синтаксического анализа.

Автоматическое обучение системы правилам синтаксического анализа осуществляется по команде подсистемы (3) управления режимом самообучения и извлечения знаний путем поиска в базе (8) стохастически индексированных лингвистических текстов фрагментов, определяющего порядок синтаксического разбора предложений. Сначала описанным выше порядком производят преобразование данных фрагментов в набор стохастически индексированных семантических структур правил продукций, имеющих в общем случае вид выражения (2).

После этого в подсистеме (6) логического вывода с использованием полученных стохастически индексированных семантических струтур (2) правил продукций, которые описывают порядок синтаксического разбора предложений, реализуется логический вывод для получения стохастически индексированных семантических струтур новых правил продукций. Эти семантические структуры связывают синтаксические элементы с заданными частями речи при формировании правил продукций, определяющих синтаксический разбор предложений по морфологическим характеристикам слов. Полученные правила заносят в базу знаний (12) синтаксического анализа, по мере заполнения которой происходит ее стохастическое индексирование и представление в виде таблицы индекса.

Как было отмечено выше, проведение синтаксического разбора текста начинается с определения порядка его реализации, который описан в учебно-методических текстовых документах по грамматике данного языка. При этом для извлечения из указанных текстов знаний, определяющих порядок синтаксического разбора, подсистемой (3) управления режимом самообучения и извлечения знаний первоначально формируется запрос к базе (8) стохастически индексированных лингвистических текстов для доступа к таблицам индексов учебно-методических текстов. По этому запросу, содержащему фразу «Порядок синтаксического разбора» на данном языке в указанных текстах будут найдены абзацы, которые включают данную фразу и термины, определяющие последовательность проведения данного разбора.

После обработки описанным выше порядком фрагмента текста, полученного из соответствующих учебно-методических изданий, для русского языка, например, может быть сформировано следующее правило продукций: «Если необходимо провести синтаксический разбор предложения, то его порядок будет следующим: словосочетание (сочинительная или подчинительная связь), простое предложение (подлежащее, сказуемое, определение, дополнение, обстоятельство), вид простого предложения (повествовательное, вопросительное, побудительное), строение предложения (двусоставное или односоставное, нераспространенное или распространенное), сказуемое (простое, составное глагольное, составное именное), предложение с однородными членами, предложение с обособленными членами, предложение с прямой речью, сложносочиненное предложение, сложноподчиненное предложение с одним придаточным, сложноподчиненное предложение с несколькими придаточными, бессоюзное сложное предложение, сложное предложение с разными видами связи». После формирования этого правила в виде выражения (2) на основе индексов $\{ I_{\xi i}^{(su)} \}$ словосочетаний формируется стохастический индекс самого правила продукций $I_{\xi i}^{(pp)}$ в виде уникальной двоичной комбинации заданной длины:

$$I_{\xi i}^{(pp)} = F\left(I_{\xi 1}^{(su)} \wedge I_{\xi 2}^{(su)} \wedge \ldots \wedge I_{\xi m}^{(su)} \Rightarrow I_{\xi 1}^{(su)} \wedge I_{\xi 2}^{(su)} \wedge \ldots \wedge I_{\xi n}^{(su)}\right), \quad (3)$$

где **F** - функция стохастического преобразования правила продукций.

Затем производится поочередное раскрытие содержания каждого из терминов, приведенных в заключение правила продукций (3), путем формирования

соответствующих запросов к базе (8) стохастически индексированных лингвистических текстов. В результате будет сформировано множество правил $\{I_{\xi ij}^{(pp)}\}$, определяющих каждый из синтаксических терминов, которые содержатся в правиле $I_{\xi i}^{(pp)}$. При этом с использованием связей между правилами продукций, включающих в условие или в заключение идентичные синтаксические термины, в подсистеме (6) реализуется логический вывод. В результате будет сформирована следующая последовательность логической связи правил продукций:

$$I_{\xi i}^{(pp)} \rightarrow \{I_{\xi i1}^{(pp)}\} \rightarrow \{I_{\xi i2}^{(pp)}\} \rightarrow \{I_{\xi i3}^{(pp)}\} \rightarrow \ldots \rightarrow \{I_{\xi i\kappa}^{(pp)}\}. \quad (4)$$

Здесь индексы $\{I_{\xi i j}^{pp)}\}$ обозначают набор правил, соответствующих определенному уровню синтаксического разбора, который задан в правиле $I_{\xi i}^{(pp)}$. Например, это может быть словосочетание (сочинительная или подчинительная связь), простое предложение (подлежащее, сказуемое, определение, дополнение, обстоятельство), вид простого предложения (повествовательное, вопросительное, побудительное) и др.

Таким образом, в системе реализуется дедуктивный логический вывод, цель которого - связать синтаксические термины с определенными частями речи слов, их характеристиками и провести последовательный синтаксический анализ согласно приведенному выше правилу. Например, для русского языка в процессе указанного логического вывода для термина «подлежащее» может быть найден следующий фрагмент текста: «Подлежащее в предложении может быть выражено следующими словами: существительным в им.п., местоимением в им.п., инфинитивом, цельным словосочетанием». Полученный фрагмент текста поступает в интерпретатор (4), подсистему (5) эквивалентных преобразований текстов и подсистему (6) логического вывода. В результате описанных выше преобразований с использованием выражения (2) получим набор правил продукций, связывающих морфологические характеристики слов с наименованиями членов предложения:

«Если в предложении есть слово, являющееся существительным в им. п., то это слово предположительно является подлежащим».

«Если в предложении есть слово, являющееся местоимением в им. п., то это слово предположительно является подлежащим».

«Если в предложении есть слово, являющееся инфинитивом, то это слово предположительно является подлежащим».

«Если в предложении есть слова, относящиеся к цельному словосочетанию, то эти слова предположительно являются подлежащим».

В процессе извлечения фрагментов текстов для формирования правил продукций, определяющих словосочетания и отдельные члены предложения, в качестве исходной информации являются морфологические характеристики слов предложения. По этим исходным данным выделяются фрагменты текста, в которых указанные данные посредством идентичных словосочетаний связаны с предполагаемым ответом, имеющим наименование члена предложения. Эти словосочетания соответствуют слову с исходными морфологическими характеристиками.

Поэтому отмеченный фрагмент текста, определяющий связь между словом с данными морфологическими характеристиками и членом предложения, может быть переведен в стохастически индексированную семантическую структуру (2) с обеспечением описанным выше порядком ее корректности. Затем стохастически индексированная семантическая структура (2) будет представлена в формате правила продукций: «Если (условие), то (заключение)». Указанная процедура осуществляется с использованием интерпретатора (4), подсистемы (5) эквивалентных преобразований текста и правил продукций. При этом в условие правила включаются исходные морфологические характеристики слова, а заключение содержит соответствующее указанному слову наименование члена предложения и вопрос, который ему соответствует.

В результате будут образованы правила продукций для определения главных членов предложения (подлежащее и сказуемое), второстепенных членов предложения (определение, дополнение, обстоятельство), а также образуемых ими словосочетаний. При определении сказуемого указывается, к какому типу оно относится: простое глагольное, составное глагольное, составное именное. Прежде всего определяется предикативная основа предложения, в котором координируют подлежащее и сказуемое, а также другие словосочетания и соответствующие им виды отношений. Они включают подлежащее и определение, сказуемое и дополнение, сказуемое и обстоятельство и т.д.

Таким образом, в процессе обработки текстовой информации при синтаксическом разборе предложения происходит самообучение системы правилам определения главных и второстепенных членов предложения. Полученные при этом правила заносятся в базу знаний (13) синтаксического анализа. Затем в соответствии с порядком синтаксического разбора начинается самообучение системы правилам определения обособленных членов предложения. Исходными данными здесь являются части речи, члены предложения и их характеристики, которые после преобразования текста входят в условия правил продукций. Заключения этих правил определяют вид группы обособленных членов, наименование члена предложения и вопрос, которым они соответствуют.

Таким образом, описывают обособленные согласованные определения (причастные обороты, прилагательные с зависимыми словами), обособленные несогласованные определения, обособленные приложения, обособленные дополнения, обособленные обстоятельства и др., включая соответствующие им вопросы.

После этого в режиме самообучения происходит формирование правил продукций, позволяющих производить разбор простого предложения на основе исходных данных, определяющих, какими членами предложений являются слова, которые входят в данное предложение, какие словосочетания и обособленные группы членов предложения они образуют. В результате будут получены правила продукций, позволяющие определить, является ли данное предложение двусоставным или односоставным (если односоставное, то к какому типу относится – неопределенно-личное, безличное, назывное и др.). При этом выделяются предложения с однородными членами, с обособленными членами предложения, с прямой речью.

Затем на основе выделяемых фрагментов текста формируются правила продукций для синтаксического разбора сложных предложений. Исходными данными, входящими в условия правил продукций, здесь являются типы и характеристики простых предложений, которые входят в состав сложных предложений. При этом заключения правил позволяют определить, к какому типу относится данное сложное предложение: сложносочиненное предложение, сложноподчиненное предложение с одним придаточным, сложноподчиненное предложение с несколькими придаточными, бессоюзное сложное предложение, сложное предложение с разными видами связей. В заключение правил также определено, какой вопрос соответствует каждому из простых предложений в составе данного сложного предложения.

Все описанные уровни формирования правил продукций соответствуют схеме разбора предложения, формируемой в начале режима самообучения по команде подсистемы (3) управления режимом самообучения и извлечения знаний в виде логического выражения (4).

В результате реализации режима самообучения полученные правила продукций записываются в базу знаний (13) синтаксического анализа. Отметим, что самообучение системы правилам синтаксического разбора предложений производится непосредственно в процессе обработки исходных текстов по заданной теме путем анализа каждого предложения. Указанный анализ позволяет заполнить уровни пять-десять фрейма каждого предложения текста, который в свою очередь используется для заполнения таблицы индексов данного текста (Таблица 2) и описанных выше списков, составляющих содержание каждой его ячейки.

По мере заполнения базы знаний синтаксического анализа происходит ее стохастическое индексирование и представление в форме таблицы индекса. Это существенно повышает эффективность разбора предложений за счет произвольного доступа по индексам условия, соответствующего правилам продукций, для получения искомого результата.

Рассмотрим более подробно порядок стохастического индексирования баз знаний и их использования в процессе грамматического разбора предложений.

После получения завершенного текста базы знаний в виде набора правил продукций, представленных в виде стохастически индексированного текста в формате «Если (условие), то (заключение)», каждое правило продукций поступает в интерпретатор (4) стохастически индексированных текстов и правил продукций. Здесь повторно формируют стохастически индексированную семантическую структуру (2), которая содержит совокупность всех словосочетаний данного правила:

$$S: \quad \{(I_{\xi i}^{(u)} I_{\xi i}^{(r)} I_{\xi i}^{(z)}) \longrightarrow (I_{\xi j}^{(u)} I_{\xi j}^{(r)} I_{\xi j}^{(z)})\} . \quad (5)$$

При этом каждому словосочетанию ставится в соответствие индекс $I_{\xi i}^{(su)}$:

$$(I_{\xi i}^{(u)} I_{\xi i}^{(r)} I_{\xi i}^{(z)}) \rightarrow (I_{\xi j}^{(u)} I_{\xi j}^{(r)} I_{\xi j}^{(z)}),$$

затем на основе этих индексов формируются уникальные стохастические индексы каждого правила продукций $I_{\xi j}^{(pp)}$ в соответствии с выражением (3).

Далее производится формирование таблицы индекса для данной базы знаний в текстовом виде подобно тому, как индексируются обычные текстовые документы. При этом в качестве абзаца принимается правило продукций с индексом $(I_{\xi j}^{(pp)})$. В соответствии с этим входом в таблицу индекса правил продукций является строка, содержащая $\{I_{\xi j}^{(u)}\}$ основ слов словаря правил продукций (множества неповторяющихся основ слов, входящих в состав правил продукций). Каждая ячейка строки, соответствующей определенному индексу $(I_{\xi j}^{(u)})$, содержит индекс $I_{\xi j}^{(su)}$ словосочетания и индекс $(I_{\xi j}^{(pp)})$ правила, который включает данное слово, окончание и номер этого слова в составе правила продукций, а также индексы $(I_{\xi j-1}^{(u)})$ и $(I_{\xi j+1}^{(u)})$, соответственно, предыдущего и последующего слова в данном правиле. Это позволяет, как и для случая с текстовыми документами, сформировать на основе индекса текст любого правила продукций. При этом выражение

$$I_{\xi i}^{(su)}: \quad (I_{\xi i}^{(u)} I_{\xi i}^{(r)} I_{\xi i}^{(z)}) \rightarrow (I_{\xi j}^{(u)} I_{\xi j}^{(r)} I_{\xi j}^{(z)})$$

записывается в виде строки таблицы базы (15) стохастически индексированных словосочетаний.

Исходные данные для обращения к индексу текста правил продукций извлекаются из фрейма разбираемого предложения. Как было представлено выше, данный фрейм после морфологического анализа содержит четыре уровня строк, включающих, соответственно, индексы основ слов $\{I_{\xi j}^{(u)}\}$, слова в контексте предложения, части речи и характеристики, соответствующие данным словам, и вопросы к ним. Именно эта информация в разных сочетаниях входит в условия правил продукций и позволяет на основе логического вывода делать заключение, к какому члену предложения (точно или неточно) относится данная часть речи. При этом обращение к таблице индексов правил продукций производится по индексам основ

слов $\{I_{\xi i}^{(u)}\}$ фрейма предложения, а также по значениям $\{I_{\xi i}^{(su)}\}$ словосочетаний условий или заключений правил.

Для реализации функций логического вывода с помощью правил продукций применяется интерпретатор (4) стохастически индексированного текста и правил продукций. В результате правило продукций преобразуется в вид (2) стохастически индексированной семантической структуры.. При этом по словосочениям $(I_{\xi i}^{(su)})$ условий правил продукций (после обращения по индексам $I_{\xi i}^{(su)}$ к базе (15) стохатически индексированных словосочетаний и определения стохастических индексов $\{I_{\xi i}^{(u)}\}$ основ слов данного словосочетания) может производиться поиск соответствующих ячеек фрейма предложения и считывание из них наименований слов, характеристик частей речи или вопросов к ним. По словосочетаниям $\{I_{\xi j}^{(su)}\}$ заключения должны заполняться соответствующие ячейки уровней 5-10 фрейма предложения, определяющих наименование членов предложения, их групп, обособленных членов, типов простых предложений в сложном предложении с указанием вопросов к ним. При этом правила продукций проверяются по всем словосочетаниям условия, и в случае истинности всех словосочетаний условия, объединенных логическими связками «и» (во фрейме предложения найдены все характеристики и данные, описанные в словосочетаниях условия правила продукций), заключение считается истинным. При этом данные, определяемые в словосочетаниях заключения правила, заносят в соответствующие ячейки фрейма предложения уровней 5-10. Если заключение содержит предварительный результат или словосочетание, по которому необходимо найти логически связанные правила, то их поиск производится путем обращения по индексам основ слов словосочетания к таблице индексов соответствующей базы знаний. При этом за счет произвольного доступа к таблицам на основе стохастических индексов исключается необходимость перебора на всем множестве правил продукций. В результате обеспечивается линейность зависимости времени логического вывода от числа задействованных в обработке правил продукций. Обращение к базе знаний и обработка правил продукций предназначены для заполнения всех ячеек фрейма предложения точными данными.

Если в процессе синтаксического анализа отдельные члены предложения будут определены неточно, то для их точного определения система переходит к семантическому анализу слов этих предложений одновременно с реализацией режима самообучения правилам семантического анализа. Это относится прежде всего к определению подлежащего, дополнения и обстоятельства, выраженным существительным с предлогом, деепричастным оборотом и др.

Для точного определения членов предложения используется семантический анализ, который основан на функции разработанной ИССИЗ, обеспечивающей выделение из текстов абзацев и предложений, описывающих все возможные виды отношений между различными объектами. Запросы системы на реализацию этой функции могут формироваться автоматически в подсистеме (3) управления режимом самообучения и извлечения знаний, если в результате синтаксического анализа не будет установлено точно, каким членом предложения являются части речи исследуемого предложения.

С этой целью используется подсистема (3) управления режимом самообучения и извлечения знаний, подсистема (6) логического вывода и интерпретатор (4) текста и правил продукций. Уточнение членов предложения в случае их неточного определения при синтаксическом анализе основано на выделении из множества текстов предложений, описывающих отношения между заданными объектами, и определении видов отношений между ними. В результате автоматического формирования запросов системы и семантического анализа выделенных предложений между заданными объектами в интерпретаторе (4) стохастически индексированного текста и правил продукций могут быть определены следующие виды отношений:

- родо-видовые,
- агрегатные (часть – целое),
- объектные отношения,
- определительные отношения,
- обстоятельственные,
- допустимые, недопустимые.

В свою очередь обстоятельственные отношения подразделяются на следующие виды:

- образа действия,
- места,

- времени,
- меры или степени,
- причины,
- цели,
- условия,
- уступки.

В тексте указанные отношения между объектами описываются предикативной основой каждого предложения, которое состоит из подлежащего и сказуемого, а также словосочетаниями между различными членами предложения и прежде всего словосочетаниями, описывающими связь сказуемого с обстоятельством (обстоятельственные отношения) или с дополнением (объектные отношения). При этом для классификации вида отношений решающую роль играют словосочетания, содержащие сказуемое и связанное с ним дополнение или обстоятельство. Именно по содержанию двух указанных членов предложения определяется, какой вид отношений имеется в данном предложении между объектами предметной области, выраженными подлежащим, а также дополнением или обстоятельством. При этом определительные отношения описывают свойства подлежащего, дополнения или обстоятельства с помощью словосочетаний, содержащих согласованные или несогласованные определения. В процессе анализа членов предложения классификация вида описываемых им отношений позволяет практически точно определить члены предложения в наиболее сложных случаях, когда синтаксический анализ дает неточный результат.

С целью классификации вида отношений в словосочетаниях в интерпретатор (4) по команде подсистемы управления (3) из таблиц индексов толковых словарей базы (8) стохастически индексированных лингвистических текстов записывают стохастические индексы типовых словосочетаний каждого из указанных выше отношений. При этом в процессе семантического анализа каждое из исследуемых словосочетаний с помощью логического вывода по таблице индексов текста толкового словаря и формирования стохастически индексированной семантической структуры соотносят с одним из индексов словосочетаний, записанных в интерпретатор (4). Порядок логического вывода по таблицам индексов текста будет представлен ниже при описании процесса формирования стохастически индексированной семантической структуры ответа системы.

В общем случае для семантического анализа слов и словосочетаний предложений в системе используется пять источников информации, а именно:

- база знаний (9), которая содержит текстовые элементы типа «запрос-ответ», формируемые в процессе функционирования ИССИЗ для обработки типовых запросов (эта база подробно будет описана ниже);

- база (8) стохастически индексированных лингвистических текстов, которая содержит таблицы индексов текстов толковых словарей, энциклопедий и базовых научно-методических материалов общего и специального назначения, позволяющих извлекать знания об объектах предметной области и видах отношений между ними;

- база знаний (14) семантического анализа, которая содержит правила для точного определения членов предложения, обеспечения эквивалентности преобразования членов предложения, которые необходимы для семантического анализа и оценки релевантности формируемых ответов на поступающие запросы; она подробно будет описана ниже;

- база знаний (12) морфологического анализа, которая содержит правила для определения частей речи и их эквивалентных преобразований;

- база знаний (13) синтаксического анализа, которая содержит правила для определения членов предложения и их эквивалентных преобразований.

Первая из названных баз знаний образуется на основе стохастически индексированных кратких ответов, формируемых в ходе обработки запросов пользователей, и содержит множество текстовых элементов типа «запрос – ответ». Эти знания представляют собой семантическую основу релевантных ответов на запросы пользователей и содержат вопросительные предложения. Каждое из данных предложений идентично соответствующему запросу пользователя, в которое после вопросительного слова (или вопросительного словосочетания) дополнительно включена соответствующая ему группа слов ответа. Эта группа может содержать одно или несколько словосочетаний, являться группой обособленных членов предложения или придаточным предложением. При этом в каждом элементе указанных знаний точно определен вопрос к группе слов ответа, что позволяет классифицировать отношения между объектами предметной области, которые представлены в данном предложении и, соответственно, определить, каким членом предложения является главное слово в словосочетании ответа.

Вторая база лингвистических текстов представлена множеством стохастически индексированных текстов, толковых словарей, энциклопедий, базовых научно-методических материалов как общего, так и тематического назначения. В их состав входит подробное описание общеупотребительной лексики, а также специальных терминов по данной теме. Эти текстовые материалы, представленные в виде таблиц индексов, используются для извлечения из них знаний, которые характеризуют базовые свойства различных типов объектов предметной области и отношения между ними, соотнося их с приведенной выше системой классификации.

Третья база знаний (14) семантического анализа состоит из правил продукций, которые сформированы автоматически и предназначены для решения задач семантического анализа текста с использованием логического вывода и информации, содержащейся в первых двух базах знаний.

Базы знаний морфологического и синтаксического анализа применяются для эквивалентных преобразований текста в ходе семантического анализа. Более подробно процесс эквивалентных преобразований будет описан ниже при анализе функций обработки запроса.

Для обеспечения рациональной обработки знаний описанная выше первая база представлена в виде таблицы индекса, вход которой включает основу слов, находящихся в знаниях «запрос - ответ». При этом каждая строка таблицы имеет ячейки, содержащие индекс текста, индекс и номер абзаца, на основе которого сформировано данное предложение, номер слова в его составе, окончание данного слова, а также индексы основ предыдущего и последующего слов в предложении. Это позволяет по запросу системы осуществлять произвольный доступ с использованием индексов основ слов к соответствующим строкам таблицы, выделять из них требуемые ячейки и при необходимости восстанавливать исходный текст соответствующего «запроса – ответа».

Описанная база знаний позволяет при синтаксическом анализе предложения определять члены предложения в наиболее сложных случаях. Например, отличить подлежащее от прямого дополнения или косвенное дополнение от обстоятельства с точной классификацией его вида и др. Для этой цели система семантического анализа формирует соответствующий запрос к базе знаний. В первом случае, когда требуется уточнить подлежащее (например, в предложениях типа ***Дождь намочил зонт*** или ***Зонт намочил дождь***), по запросу системы определяют, для какого объекта является

допустимым отношение, выраженное сказуемым. При этом очевидно, что объект, соответствующий допустимому отношению, принимается в качестве подлежащего.

В случае, когда база знаний не позволяет дать ответ на указанный запрос, вопрос будет обращен к таблицам индексов текстов по данной проблематике для поиска словосочетания, содержащего требуемое отношение между объектами на всем множестве текстовых документов второй базы знаний по данной теме.

Во втором случае на основе запроса системы к базе знаний должно быть определено, на какой вопрос отвечает член предложения, который можно отнести как к дополнению, так и к обстоятельству и тем самым точно установить, каким членом предложения является данное слово. Для этой цели в запросе системы, обращенном к базе знаний, указывается требуемое слово и предполагаемый вопрос к нему. Если при этом в базе знаний находится соответствующий «запрос – ответ», у которого в словосочетании ответа главное слово и вопрос к нему совпадают, соответственно, с содержанием запроса системы, то это означает, что анализируемый член предложения точно отвечает на данный вопрос. Следовательно, указанный результат обработки запроса системы позволяет точно определить, каким членом предложения является содержащееся в нем слово. Например, если анализируется предложение типа ***Мужчина прогуливается в парке*** или ***Мужчина прогуливается в костюме*** для уточнения, каким членом предложения (обстоятельством или дополнением) являются словосочетания ***в парке*** или ***в костюме***, формируется два запроса системы. Первый запрос содержит вопросительное слово ***где?*** и словосочетание ***в парке***, поскольку в результате синтаксического анализа был сделан неточный вывод о том, что ***в парке*** – это обстоятельство места. Во втором случае формируется следующий запрос системы: ***в чем? – в костюме***. Если в результате обработки запроса системы будет дан положительный ответ на каждый из них, то это означает, что первое словосочетание является точно обстоятельством, а второе – дополнением. Если будет сформирован запрос системы, содержащий ошибочное утверждение (например, ***где? – в костюме***), то ответ будет отрицательным. Это означает, что словосочетание ***в костюме*** не является обстоятельством места.

Описанный способ формирования запросов к первой базе знаний системы семантического анализа может быть использован и в более сложных случаях синтаксического анализа предложений. Например, при определении, каким видом обстоятельства является деепричастный оборот (деепричастие), или при уточнении

типа придаточного предложения. Для этой цели формируется специальный запрос, содержащий данный деепричастный оборот или придаточное предложение, на основе которого с точностью до синонимов производится поиск их аналогов на множестве знаний типа «запрос-ответ». Если указанные аналоги содержатся в группе слов ответа этой базы, то с использованием индексной таблицы текста они будут извлечены из нее. Это позволит определить вопрос, которому соответствует определяемый деепричастный оборот или придаточное предложение и, следовательно, точно выявить, к какому типу они относятся.

Если в первой базе знаний не содержится запрашиваемых аналогов, то для точного определения членов предложения используется вторая и третья базы знаний в сочетании с подсистемой (6) логического вывода. Как было отмечено выше, третья база знаний составлена из правил продукций, которые позволяют с помощью семантического анализа уточнять наименования членов предложения, деепричастных оборотов или типов придаточных предложений в сложноподчиненных предложениях с целью формирования к ним соответствующих вопросов.

Одним из основных вариантов проведения семантического анализа с использованием этой базы знаний является перевод с помощью правил продукций семантических определений, характерных для каждого члена предложения, в набор словосочетаний, содержащих определяемое слово и некое базовое слово. Это базовое слово семантически связано только с данным членом предложения и однозначно ему соответствует (не может употребляться с другими членами предложения). При формировании из исходного анализируемого текста словосочетания, описанного в правилах продукций, часто необходимо проводить эквивалентные преобразования исходного текста на основе правил баз знаний морфологического, синтаксического анализа с использованием логического вывода.

После получения требуемого словосочетания проводится проверка его допустимости путем обращения ко второй индексированной базе текстов, которая позволяет производить выделение абзацев и отдельных предложений, содержащих требуемые словосочетания. Если на множестве текстовых документов найдется одно или более предложений, в которых данное словосочетание используется, то отношения между словами данного словосочетания являются допустимыми. Поэтому считается, что исследуемое слово точно относится к данному члену предложения.

Вместо отдельных словосочетаний могут использоваться более сложные конструкции (например, причастный, деепричастный обороты, придаточные предложения в сложных предложениях). Таким образом, сочетание семантических знаний, выраженных конкретными словосочетаниями, в совокупности с определением допустимости отношений между словами в них на множестве текстовых документов позволит точно определять члены предложения, если их синтаксический анализ не дает точный результат.

После завершения морфологического, синтаксического и семантического анализа предложений данного текстового документа на основе полученных при этом фреймов предложений полностью заполняется таблица индексов данного текста (Таблица 2), включая списки $\{I_{ij}^{(s)}\}$, определяющие содержание каждой ячейки таблицы. После этого переходят к стохастическому индексированию следующего текста по данной теме. Одновременно с этим реализуется автоматическое обучение и происходит заполнение базы знаний (14) семантического анализа правилами продукций, сформированными на основе соответствующих фрагментов текста описанным выше порядком с использованием стохастически индексированной семантической структуры (2). Отметим, что корректность каждого правила обеспечивается при этом путем независимого формирования описанным выше порядком идентичных стохастически индексированных семантических структур (2) на основе нескольких фрагментов из соответствующих лингвистических текстов. Затем стохастически индексированная семантическая структура переводится в формат правил продукций, представленный в виде «Если (условие), то (заключение)». Это происходит в интерпретаторе (4) стохастически индексированных текстов и правил продукций и в подсистеме (5) эквивалентных преобразований текста.

После обработки всех представленных текстовых документов по данной теме формируется таблица индексов текстов по данной теме (Таблица 3). Ее строки поименованы неповторяющимися индексами $\{I_{ij}^{(u)}\}$ основ слов, входящих в текстовые документы. Столбцы данной таблицы соответствуют стохастическим индексам $\{I_{ij}^{(t)}\}$ текстов, которые были обработаны в ходе грамматического и семантического анализа. Ячейки этой таблицы содержат индексы $\{I_{ij}^{(s)}\}$ списков, содержащих индексы абзацев

$\{I_{\xi i}^{(a)}\}$ каждого текста $I_{\xi j}^{(t)}$, в которые входит соответствующий индекс $I_{\xi j}^{(u)}$ основы слова. Записи списков хранятся в отдельном файле, доступ к которым производится по соответствующим индексам $\{I_{\xi i}^{(s)}\}$.

После формирования указанных таблиц индексов и заполнения баз знаний в режиме самообучения ИССИЗ по команде подсистемы (3) управления режимом самообучения и извлечения знаний переходят к обработке запроса пользователя с целью извлечения знаний из текстовых документов, релевантных этому запросу.

В данном процессе широко используются эквивалентные преобразования как запроса пользователя, так и предложений фрагментов текста при извлечении из них знаний. Рассмотрим более подробно порядок преобразований предложений текста.

В ИССИЗ обеспечивают следующие уровни эквивалентных преобразований текста.

Первый уровень преобразований реализуется внутри групп членов предложений – словосочетаний, содержащих подлежащее, сказуемое, дополнение, обстоятельство. При этом происходит изменение частей речи с целью замены согласованных определений на несогласованные. Этому уровню соответствуют преобразования терминов, например: ***компьютерная сеть – сеть компьютеров, компьютерное обслуживание – обслуживание компьютеров.***

Второму уровню преобразований соответствуют эквивалентные преобразования членов предложения внутри простых предложений как самостоятельных, так и составляющих сложные. При этом реализуются следующие виды замены членов предложения с использованием преобразований однокоренных частей речи:

подлежащее заменяется на сказуемое,

сказуемое – на подлежащее,

дополнение – на подлежащее,

сказуемое – на обстоятельство и т.д.

В частных случаях части речи могут не изменяться (изменяются только падежи).

Третий уровень эквивалентных преобразований соответствует преобразованию внутри сложных предложений. В этом случае придаточное предложение одного вида может быть заменено на придаточное предложение другого вида или на причастные, деепричастные обороты. Иногда сложное предложение преобразуется в простое

предложение путем замены союза на соответствующие предлоги, определяемые правилами.

Рассмотрим пример эквивалентных преобразований с использованием замены членов предложения в словосочетаниях, а именно: замены согласованного определения на несогласованное и прямого дополнения на подлежащее. Выберем в качестве исходного предложения следующее: **«Программные и аппаратные средства защищают компьютерные программы».** В системе исходное предложение с индексом $I_{\xi 1}^{(p)}$ будет представлено приведенной ниже стохастически индексированной семантической структурой:

$$I_{\xi 1}^{(p)} : \quad I_{\xi 12}^{(su)} \rightarrow I_{\xi 13}^{(su)} \rightarrow I_{\xi 14}^{(su)}. \quad (5а)$$

Эта структура содержит следующие словосочетания исходного предложения:

$I_{\xi 12}^{(su)}$: = (программные и аппаратные средства),

$I_{\xi 13}^{(su)}$: = (защищают),

$I_{\xi 14}^{(su))}$: = (компьютерные программы).

Произведем указанные выше эквивалентные преобразования членов предложения. При этом будут образованы такие словосочетания:

$I_{\xi 22}^{(su)}$: = (программы компьютера),

$I_{\xi 23}^{(su)}$: = (защищаются),

$I_{\xi 24}^{(su)}$: = (программными и аппаратными средствами).

В результате данных преобразований будет получено предложение, эквивалентное исходному предложению с индексом $I_{\xi 1}^{(p)}$, которое имеет индекс $I_{\xi 2}^{(p)}$ и следующую стохастически индексированную семантическую структуру:

$$I_{\xi 2}^{(p)} : \quad I_{\xi 22}^{(su)} \rightarrow I_{\xi 23}^{(su)} \rightarrow I_{\xi 24}^{(su)}. \quad (5б)$$

На основе этой структуры будет образовано предложение: **«Программы компьютера защищаются программными и аппаратными средствами»**, которое эквивалентно исходному. Отметим, что в новом предложении подлежащее $I_{\xi 22}^{(su)}$

соответствует словосочетанию прямого дополнения $I_{\xi\ 14}^{(su)}$ исходного предложения, в котором произведена замена согласованного определения на несогласованное. При этом подлежащее первого предложения $I_{\xi 12}^{(su)}$ преобразовано в косвенное дополнение $I_{\xi 24}^{(su)}$ во втором предложении, а сказуемое $I_{\xi 13}^{(su)}$ стало иметь форму возвратного глагола $I_{\xi\ 23}^{(su)}$. Указанные преобразования наиболее часто используются как для эквивалентных преобразований стохастически индексированных предложений текста, так и для запросов пользователей.

Запрос пользователя формируют на естественном языке. Затем преобразуют запрос пользователя во множество новых запросов, включающих вопросительное слово и словосочетания, определяющие семантику запроса, эквивалентных исходному запросу. Указанные эквивалентные преобразования исходного запроса пользователя осуществляют с использованием синонимов, близких по смыслу слов, а также замены частей речи и членов предложения. При этом обеспечивается сохранение смыслового содержания исходного запроса на основе применения стохастически индексированных правил морфологического, синтаксического и семантического анализа для получения эквивалентных структур словосочетаний вопросительного предложения запроса и сохранения семантической связи между ними.

После этого в соответствии с очередным преобразованным запросом пользователя осуществляют предварительный выбор фрагментов текстовых документов, содержащих в совокупности все словосочетания запроса. Если данный запрос не обеспечил возможность предварительного выбора фрагментов текстовых документов, отвечающих указанным требованиям, то производят новое эквивалентное преобразование запроса.

Рассмотрим порядок обработки запроса и алгоритма формирования ответа на основе различных текстовых документов, абзацев и предложений. После поступления очередного запроса пользователя в лингвистический процессор (1) он заносится в подсистему (2) стохастического индексирования и выделения фрагментов текстов, где производится формирование стохастических индексов основ слов и выделение их окончаний. После этого стохастически индексированный запрос через подсистему (3) управления режимом самообучения и извлечения знаний записывается в подсистему

(6) логического вывода. Здесь на основе правил продукций баз знаний (12-13) сначала производят морфологический и синтаксический разбор запроса пользователя. Получают фрейм вопросительного предложения. Затем в интерпретаторе (4) вопросительное предложение представляют в виде совокупности словосочетаний, содержащих главные и зависимые слова, и соответствующих им стохастических индексов основ слов

$$S: \{(I_{\xi i}^{(u)} I_{\xi i}^{(r)} I_{\xi i}^{(z)}) \longrightarrow (I_{\xi j}^{(u)} I_{\xi j}^{(r)} I_{\xi j}^{(z)})\}, \qquad (6)$$

где $I_{\xi i}^{(u)} I_{\xi j}^{(u)}$ - стохастические индексы соответственно главного и зависимого основ слов данного словосочетания,

$I_{\xi i}^{(r)} I_{\xi j}^{(r)}$ - стохастические индексы частей речи главного и зависимого слов данного словосочетания,

$I_{\xi i}^{(z)} I_{\xi j}^{(z)}$ - стохастические индексы соответственно морфологических и синтаксических характеристик частей речи главного и зависимого слов данного словосочетания.

На основе полученных индексов формируют стохаситчески индексированную семантическую структуры запроса, которая в общем случае имеет следующий вид:

$$P: I_{\xi 1}^{(su)} \wedge I_{\xi 2}^{(su)} \longrightarrow I_{\xi 3}^{(su)} \longrightarrow I_{\xi 4}^{(su)} \wedge I_{\xi 5}^{(su)}, \qquad (7)$$

где $I_{\xi 1}^{(su)}$ – индекс вопросительного словосочетания,

$I_{\xi 2}^{(su)}$ – индекс словосочетания подлежащего, $I_{\xi 3}^{(su)}$ - индекс словосочетания сказуемого, $I_{\xi 2}^{(su)} \longrightarrow I_{\xi 3}^{(su)}$ - предикативная основа предложения, связывающая подлежащее и сказуемое, $I_{\xi 3}^{(su)} \longrightarrow I_{\xi 4}^{(su)}$ связь между сказуемым и дополнением (обстоятельством), определяющая тип отношения в данном предложении, $I_{\xi 4}^{(su)}$ - индекс словосочетаний дополнения (обстоятельства), $I_{\xi 5}^{(su)}$ - индекс словосочетания обстоятельства (дополнения).

По полученным индексам выражений (6,7) путем обращения к базе (10) стохастически индексированных текстов по заданным темам с использованием таблиц индексов текстов по заданной теме (Фиг. 4) находят совокупность фрагментов, в которые входят все словосочетания запроса, включая вопросительное словосочетание. При этом каждый фрагмент текста может состоять из одного или нескольких абзацев.

Если будут найдены один или несколько текстов, отвечающих указанным условиям, то переходят к дальнейшей обработке абзацев этих текстов с использованием таблиц индексов каждого из них. Отметим: наличие в таблице индексов одного из текстов индекса $I_{\xi\ 1}^{(su)}$ вопросительного словосочетания, содержащего индекс вопроса (в списке $I_{\xi j}^{(s)}$ одной из ячеек таблицы) и связанного с ним индекса основы главного слова, свидетельствует о том, что в указанном абзаце данного текста есть предложение, которое содержит группу слов ответа $I_{\xi\ 0}^{(su)}$, связанную с главным словом вопросительного словосочетания: $(I_{\xi 0}^{(su)} \rightarrow I_{\xi 1}^{(su)})$.

Если не будет найден хотя бы один из текстов, отвечающий данным условиям, то переходят к эквивалентным преобразованиям запроса пользователя путем замены слов, которые не вошли в абзац текста, на синонимы и близкие по смыслу слова, а также применяя замену частей речи и членов предложения без изменения смысла запроса.

Дальнейшую обработку текста, отвечающего указанным выше условиям, производят по таблице индексов данного текста. С этой целью, используя индексы вопросительного словосочетания $I_{\xi 1}^{(su)}$ путем обращения к таблице индексов текста из базы (10), находят предложение, содержащее группу слов ответа, которая соответствует вопросительному словосочетанию запроса и связана с главным словом этого запроса. Если словосочетания

$$S : \{(I_{\xi i}^{(u)} I_{\xi i}^{(r)} I_{\xi i}^{(z)}) \rightarrow (I_{\xi j}^{(u)} I_{\xi j}^{(r)} I_{\xi j}^{(z)})\}$$

запроса при этом входят в разные абзацы различных текстов V: $\{(I_{\xi i}^{(t)} I_{\xi j}^{(a)})\}$, то необходимым условием для формирования единого, логически связанного текста ответа является наличие хотя бы в одном из абзацев группы слов ответа $I_{\xi 0}^{(su)}$,

соответствующего $I_{\xi 1}^{(su)}$ вопросительного словосочетания запроса, и предикативной основы $I_{\xi 2}^{(su)} \longrightarrow I_{\xi 3}^{(su)}$ выражения (7), в которую в общем виде входят индексы словосочетаний подлежащего и сказуемого. Если указанное условие выполняется, то выделенная совокупность абзацев используется при дальнейшей обработке, поскольку на основе предварительно выбранных абзацев можно попытаться сформировать единый, логически связанный текст ответа. В противоположном случае необходимо перейти к вводу и индексированию новых текстов по данной теме.

Рассмотрим сначала более простой случай формирования релевантного ответа, когда фрагмент текста, содержащего все словосочетания запроса, может быть образован на основе одного или нескольких последовательных абзацев данного текста. В этом случае сначала формируют основу стохастически индексированной семантической структуры ответа пользователя в виде следующего выражения:

$$P: I_{\xi 0}^{(su)} \longrightarrow I_{\xi 1}^{(su)} \wedge I_{\xi 2}^{(su)} \longrightarrow I_{\xi 3}^{(su)}, \qquad (8)$$

где $I_{\xi 0}^{(su)}$ - индекс группы слов ответа, $I_{\xi 1}^{(su)}$ - индекс вопросительного словосочетания, $I_{\xi 2}^{(su)}$ - индекс словосочетания подлежащего, $I_{\xi 3}^{(su)}$ - индекс словосочетания сказуемого, $I_{\xi 2}^{(su)} \longrightarrow I_{\xi 3}^{(su)}$ - предикативная основа предложения. С этой целью после определения в данном фрагменте текста предложения, где в индексированном виде содержится группа слов ответа, связанная с главным словом вопросительного словосочетания ($I_{\xi 0}^{(su)} \longrightarrow I_{\xi 1}^{(su)}$), находят предложение, в которое входит предикативная основа ($I_{\xi 2}^{(su)} \longrightarrow I_{\xi 3}^{(su)}$).

Поскольку указанные группы слов в общем случае входят в разные выражения, то для образования семантической структурной схемы (8) реализуют процедуру логического вывода с использованием индексированных предложений данного фрагмента текста. С этой целью предложение с номером i, содержащее группу слов ответа, представляют в следующем виде:

$$P: I_{\xi 0}^{(su)} \wedge I_{\xi 1}^{(su)} \wedge I_{\xi 2 i}^{(su)} \longrightarrow I_{\xi 3 i}^{(su)} \longrightarrow I_{\xi 4 i}^{(su)} \wedge I_{\xi 5 i}^{(su)} \qquad (9)$$

где $I_{\xi 0}^{(su)}$ - индекс группы слов ответа, $I_{\xi 1}^{(su)}$ - индекс вопросительного словосочетания, $I_{\xi 2 i}^{(su)}$ - индекс словосочетания подлежащего, $I_{\xi 3 i}^{(su)}$ - индекс словосочетания сказуемого, $I_{\xi 2 i}^{(su)} \longrightarrow I_{\xi 3 i}^{(su)}$ - предикативная основа предложения , $I_{\xi 3 i}^{(su)} \longrightarrow I_{\xi 4 i}^{(su)}$ - связь между сказуемым и дополнением (обстоятельством), определяющая тип отношения в данном предложении, $I_{\xi 4 i}^{(su)}$ - индекс словосочетаний дополнения (обстоятельства), $I_{\xi 5 i}^{(su)}$ - индекс словосочетания обстоятельства (дополнения).

Для реализации логического вывода на основе выражения (9) с использованием транзитивной зависимости формируется стохастически индексированная семантическая структура типа **тема** $\longrightarrow$ **рема** предложения с номером **i**:

$$TR : I_{\xi 2i}^{(su)} \longrightarrow I_{\xi 3 i}^{(su)} \longrightarrow I_{\xi 4 i}^{(su)} = I_{\xi 2i}^{(su)} \longrightarrow I_{\xi 4 i}^{(su)} \qquad (10)$$

где тема является индексом $I_{\xi 2i}^{(su)}$ словосочетания подлежащего, а рема - индексом $I_{\xi 4i}^{(su)}$ словосочетания дополнения (обстоятельства).

При этом предложение с номером **j**, содержащее предикативную основу запроса, имеет в общем случае следующую стохастически индексированную семантическую структуру:

$$P : I_{\xi 2}^{(su)} \longrightarrow I_{\xi 3}^{(su)} \longrightarrow I_{\xi 4 j}^{(su)} \wedge I_{\xi 5 j}^{(su)} \qquad (11)$$

где $I_{\xi 2}^{(su)}$ - индекс словосочетания подлежащего запроса, $I_{\xi 3}^{(su)}$ - индекс словосочетания сказуемого запроса, $I_{\xi 2}^{(su)} \longrightarrow I_{\xi 3}^{(su)}$ - предикативная основа предложения запроса, $I_{\xi 3}^{(su)} \longrightarrow I_{\xi 4 j}^{(su)}$ - связь между сказуемым и дополнением (обстоятельством), определяющая тип отношения в данном предложени с номером **j**, $I_{\xi 4}^{(su)}$ - индекс словосочетаний дополнения (обстоятельства), $I_{\xi 5 j}^{(su)}$ - индекс словосочетания обстоятельства (дополнения). Затем выражение (11) преобразуется в следующую семантическую структуру **тема** $\longrightarrow$ **рема** предложения с номером **j**:

$$TR: \quad I_{\xi 2}^{(su)} \longrightarrow I_{\xi 4j}^{(su)} \qquad (12)$$

Отметим, что в текстовой информации между законченными предложениями имеется семантическая, а следовательно, и грамматическая (синтаксическая) связь. Существуют два способа структурной соотнесенности предложений - синтаксической связи между ними. Первый способ можно назвать цепной (последовательной), а второй – параллельной связью.

Цепная связь отражает последовательное развитие мысли в связном тексте. Тема - это исходный пункт, начало движения мысли, «данное», рема - развитие мысли, ее основа, ядро, «новое».

Синтаксический характер цепной связи выражается в структурной соотнесенности двух соседних предложений. Обычно какой-либо член предшествующего предложения, например, дополнение, в последующем предложении становится подлежащим. Наиболее распространенные структурные виды цепной связи «дополнение – подлежащее», «дополнение – дополнение», «подлежащее – дополнение», «подлежащее – подлежащее» и др.

Структурная соотнесенность между предложениями при цепной связи выражается: а) с помощью лексического повтора (когда соотносящиеся члены предложений выражены одинаково); б) посредством синонимической лексики; в) с помощью местоимений.

Цепная связь - один из важнейших и наиболее распространенных способов связи самостоятельных предложений.

Параллельная связь, как и цепная, заключается в структурной соотнесенности соединяемых предложений. Однако характер этой соотнесенности иной. Основные структурные признаки параллельной связи предложений: а) параллелизм структуры (однотипность или синтаксическая близость соединяемых предложений); б) параллельный (сходный) порядок слов; в) одинаковое грамматическое выражение всех или некоторых членов предложений.

Семантическим «входом» как в цепную, так и в параллельную структуры связи абзаца является тема начального ее предложения в связанных предложениях данного абзаца или нескольких последовательных абзацев текста.

В соответствии с этим на основе элементарной семантической структуры каждого предложения типа **тема→ рема** с помощью логического вывода могут быть сформированы более сложные семантические структуры, определяющие связи между предложениями как последовательного, так и параллельного типа. Поэтому необходимым условием семантической связи между группой слов ответа, содержащейся в предложении с номером **i** , и предикативной основой предложения запроса, которая входит в состав предложения с номером **j**, является доказательство с помощью логического вывода их вхождения в единую семантическую структуру данного фрагмента текста. В стохастически индексированном виде эта структура может выглядеть следующим образом:

$$I_{\xi 0}^{(su)} \wedge I_{\xi 1}^{(su)} \wedge I_{\xi 2i}^{(su)} \rightarrow I_{\xi 4i}^{(su)} \wedge I_{\xi 4i}^{(su)} \rightarrow I_{\xi 4k}^{(su)} \wedge \ldots \wedge I_{\xi 2m}^{(su)} \rightarrow I_{\xi 2}^{(su)} \wedge \quad (13)$$

$$I_{\xi 2}^{(su)} \rightarrow I_{\xi 4j}^{(su)} = I_{\xi 0}^{(su)} \wedge I_{\xi 1}^{(su)} \wedge I_{\xi 2i}^{(su)} \rightarrow I_{\xi 4i}^{(su)} \rightarrow I_{\xi 4k}^{(su)} \ldots I_{\xi 2}^{(su)} \rightarrow I_{\xi 4j}^{(su)}$$

Логический вывод для установления семантической связи между указанными группами слов производят по таблице индексов текста базы (10) стохастически индексированных текстовых документов по заданным темам. С этой целью используется подсистема (6) логического вывода и подсистема (5) эквивалентных преобразований текста. При этом логический вывод начинается с предложения с номером **i**, содержащего группу слов ответа, которая связана с главным словом вопросительного словосочетания, предикативную основу запроса, и имеет стохастически индексированную семантическую структуру (9).

После представления названного предложения в виде семантической структуры типа **тема → рема** (10) по таблице индексов находят следующее предложение, в котором рема данного предложения переходит в тему следующего предложения. Для этого используют ячейки, которые соответствуют индексу данного абзаца $I_{\xi j}^{(a)}$ и индексу словосочетания $I_{\xi 4i}^{(su)}$, являющегося дополнением или обстоятельством предложения с номером **i**. По этим ячейкам находят номер предложения данного абзаца, в котором данное словосочетание включает подлежащее. Затем, используя адресную информацию ячейки, находят индекс сказуемого указанного предложения и связанные с ним индексы

словосочетания дополнения или обстоятельства $I_{\xi\ 4\ k}^{(su)}$, т.е. в соответствии с выражением (13) рему следующего предложения, логически связанного с предыдущим и т.д. Логический вывод продолжается до тех пор, пока в очередном предложении, определяемом связью $(I_{\xi 2}^{(su)} \rightarrow I_{\xi\ 4j}^{(su)})$, не будут содержаться индексы $(I_{\xi 2}^{(su)} \rightarrow I_{\xi\ 3}^{(su)})$, которые соответствуют предикативной основе запроса.

Если в ходе логического вывода индекс ремы $I_{\xi\ 4n}^{(su)}$ очередного предложения не совпадает с темой $I_{\xi 2n+1}^{(su)}$ последующего предложения, то это означает, что в последующем предложении используется либо синоним данного слова, либо местоимение. В первом случае по индексам основ слов $I_{\xi 2n+1}^{(su)}$ этого словосочетания обращаются к таблице индексов словаря синонимов базы (8) стохастически индексированных лингвистических текстов. Здесь находят основы слов синонимов $\{I_{\xi s}^{(u)}\}$, из которых можно образовать индекс $I_{\xi\ 4\ n}^{(su)}$ ремы предыдущего предложения. Во втором случае индекс $I_{\xi 2n+1}^{(su)}$ темы следующего предложения может соответствовать местоимению, согласованному со словосочетанием $I_{\xi 4n}^{(su)}$, что проверяется по таблице индексов словаря базы данных (7). При выполнении первого или второго условия логический вывод продолжается, пока не будет найдено предложение, содержащее искомое словосочетание запроса, в данном случае $(I_{\xi\ 2}^{(su)} \rightarrow I_{\xi 3}^{(su)})$ предикативной основы запроса. Таким образом, в ходе логического вывода будет синтезирована стохастически индексированная семантическая структура, описанная выражением (13).

Поскольку в рассмативаемом случае все словосочетания запроса входят в один абзац или в группу последовательных абзацев одного текста, то логический вывод в данном фрагменте текста будут продолжать с целью образования единой стохастически индексированной семантической структуры, содержащей все словосочетания запроса, включая словосочетания дополнения $I_{\xi\ 4}^{(su)}$ и обстоятельства $I_{\xi\ 5}^{(su)}$:

$$S: \quad I_{\xi 0}^{(su)} \wedge I_{\xi 1}^{(su)} \wedge I_{\xi 2i}^{(su)} \rightarrow I_{\xi\ 4i}^{(su)} \rightarrow I_{\xi\ 4k}^{(su)} \ldots I_{\xi 2}^{(su)} \rightarrow$$

$$I_{\xi 4}^{(su)} \dots I_{\xi 2m}^{(su)} \rightarrow I_{\xi 4}^{(su)} \dots I_{\xi 2n}^{(su)} \rightarrow I_{\xi 5}^{(su)} \qquad (14).$$

С этой целью реализуют описанные выше функции логического вывода по схеме **тема → рема** до тех пор, пока все словосочетания запроса, входящие в различные предложения данного абзаца, будут включены в семантическую структуру (14). Отметим, что необходимым условием синтеза указанной семантической структуры (14) является соответствие словосочетаний запроса и идентичных им словосочетаний в тексте абзаца одним и тем же членам предложений. Поэтому, если некоторые словосочетания, идентичные словосочетаниям запроса в предложениях текста, относятся к другим членам предложения, то эти предложения подвергают эквивалентным преобразованиям с тем, чтобы указанные словосочетания относились к требуемым членам предложений. Эти функции выполняют описанным выше порядком в подсистеме (5) эквивалентных преобразований текста.

После образования семантической структуры (14) переходят к контролю ее непротиворечивости. С этой целью проверяют семантическое соответствие словосочетания сказуемых $\{I_{\xi 3i}^{(su)}\}$, входящих в каждое из предложений, на основе которых образована семантическая структура (14), базовым отношениям. К ним относятся родо-видовые отношения, отношения типа “часть – целое” или “причина-следствие” (условие-заключение). Эти отношения определяются путем обращения по указанным индексам к базе (8) стохастически индексированных текстов для поиска семантических значений сказуемых $\{I_{\xi 3i}^{(su)}\}$ в таблицах индексов толковых словарей. При этом проверяется идентичность семантических значений сказуемых $\{I_{\xi 3i}^{(su)}\}$ индексам указанных выше базовых отношений или их синонимов, записанных в интерпретатор (4). В случае выполнения данных условий в образованной семантической структуре (14) поддерживается транзитивная зависимость. Поэтому любое искомое словосочетание запроса с индексом $I_{\xi j}^{(su)}$ может быть перенесено в формируемое предложение ответа с использованием логического вывода на образованной семантической структуре типа **тема→ рема** после словосочетания с индексом $I_{\xi j-1}^{(su)}$. Если это условие не выполняется, то данный абзац не содержит ответа, релевантного

запросу пользователя. В этом случае переходят к анализу следующего предварительно выбранного абзаца или совокупности последовательных абзацев.

Описанную процедуру логического вывода для определения семантической связи между словосочетаниями запроса при нахождении их в различных предложениях абзаца производят до тех пор, пока не будет сформирован краткий ответ пользователю в виде предложения, содержащего группу слов ответа, вопросительное словосочетание, предикативную основу и все другие словосочетания, которые входят в ответ. При этом сформированный краткий ответ будет представлен в виде следующей стохастически индексированной семантической структуры:

$$P: I_{\xi 0}^{(su)} \wedge I_{\xi 1}^{(su)} \wedge I_{\xi 2}^{(su)} \longrightarrow I_{\xi 3}^{(su)} \longrightarrow I_{\xi 4}^{(su)} \wedge I_{\xi 5}^{(su)} \qquad (15)$$

где $I_{\xi 0}^{(su)}$ - индекс группы слов ответа, $I_{\xi 1}^{(su)}$ - индекс вопросительного словосочетания, $I_{\xi 2}^{(su)}$ - индекс словосочетания подлежащего, $I_{\xi 3}^{(su)}$ - индекс словосочетания сказуемого, $I_{\xi 2}^{(su)} \longrightarrow I_{\xi 3}^{(su)}$ - предикативная основа предложения, $I_{\xi 3}^{(su)} \longrightarrow I_{\xi 4}^{(su)}$ - связь между сказуемым и дополнением (обстоятельством), определяющая тип отношения в данном предложени, $I_{\xi 4}^{(su)}$ - индекс словосочетаний дополнения (обстоятельства), $I_{\xi 5}^{(su)}$ - индекс словосочетания обстоятельства (дополнения).

При этом корректность краткого ответа обеспечивают путем формирования описанным выше порядком нескольких идентичных стохастически индексированных семантических структур (15) на основе различных, предварительно выбранных стохастически индексированных фрагментов текстовых документов.

Сформированное выражение (15) означает, что в результате логического вывода получен краткий ответ, идентичный вопросительному предложению запроса. Поэтому данный ответ является релевантным запросу пользователя. Он может быть выдан пользователю после преобразования в текстовую форму на данном языке в виде знания, сформированного системой в соответствии с его запросом.

При необходимости получения по требованию пользователя более полного ответа переходят к преобразованию исходного абзаца текста, на основе которого сформирован краткий ответ, а при необходимости и последующих абзацев текста. Это производят с целью получения на основе указанных абзацев единой стохастически индексированной семантической структуры, дающей возможное уточнение краткого ответа в рамках данного фрагмента текста. Описанные функции формирования полного ответа будут представлены ниже.

Если же в результате предварительного поиска по таблице индексов текстов не будут найдены тексты, содержащие абзацы, включающие все словосочетания ответа, то по полученным индексам запроса находят тексты, фрагменты которых в совокупности включают все словосочетания запроса. Если такая совокупность не будет найдена, то это означает, что содержание базы (10) стохастически индексированных текстовых документов не позволяет сформировать ответ, релевантный запросу пользователя. В этом случае необходимо перейти к вводу и индексированию новых текстов по данной теме из поисковой системы.

В процессе предварительного выбора, используя таблицу индексов текстов по индексам словосочетаний $S: \{ I_{\xi i}^{(u)} \rightarrow I_{\xi j}^{(u)} \}$ запроса, выбирают для каждого текста фрагменты в виде совокупности абзацев, содержащих все словосочетания запроса

$$V: = \{I_{\xi i}^{(t)}, I_{\xi j}^{(a)}\},$$

где $I_{\xi i}^{(t)}, I_{\xi j}^{(a)}$ - соответственно индекс текста и индекс абзаца данного текста, содержащих определенные словосочетания запроса пользователя. Если индексы $I_{\xi i}^{(su)}: \{I_{\xi i}^{(u)} \rightarrow I_{\xi j}^{(u)}\}$ словосочетаний запроса не входят в полном составе ни в один абзац $(I_{\xi i}^{(t)}, I_{\xi j}^{(a)})$ хотя бы одного из текстов $I_{\xi i}^{(t)}$, а содержатся в различных абзацах одного текста или в различных абзацах разных текстов $V: = \{I_{\xi i}^{(t)}, I_{\xi j}^{(a)}\}$, то на основе предварительно выбранных абзацев фрагментов текстов необходимо сформировать единый логически связанный текст, содержащий все словосочетания запроса

$S: = \{I_{\xi i}^{(su)}\}$, включая вопросительное словосочетание.

Если словосочетания S: = $\{I_{\xi i}^{(su)}\}$ при этом входят в разные абзацы различных текстов V: = $\{I_{\xi i}^{(t)}, I_{\xi j}^{(a)}\}$, то необходимым условием для формирования единого, логически связанного текста ответа является наличие хотя бы в одном из абзацев индексов группы слов ответа $I_{\xi 0}^{(su)}$, главного слова вопросительного словосочетания $I_{\xi 1}^{(su)}$ запроса и предикативной основы $(I_{\xi 2}^{(su)} \rightarrow I_{\xi 3}^{(su)})$ выражения (15), в которую в общем виде входят индексы словосочетаний подлежащего и сказуемого.

Если указанное условие выполняется, то выделенная совокупность абзацев используется при дальнейшей обработке, поскольку на основе предварительно выбранных абзацев можно попытаться сформировать единый, логически связанный текст ответа. В противоположном случае необходимо перейти к вводу и индексированию новых текстов по данной теме.

При выполнении указанного условия переходят к формированию логически связанной совокупности указанных абзацев. С этой целью проверяют выполнение следующего условия: каждое словосочетание входит не менее, чем в два различных абзаца:

$$I_{\xi i}^{(su)} \rightarrow (I_{\xi i}^{(t)}, I_{\xi j}^{(a)}), \ldots, (I_{\xi k}^{(t)}, I_{\xi l}^{(a)}). \qquad (16)$$

При невыполнении этого условия проверяется, есть ли в абзацах, содержащих только одно словосочетание запроса $I_{\xi i}^{(su)}$, другое словосочетание $I_{\xi k}^{(su)}$, которое содержится в других предварительно выбранных абзацах и связано со словосочетанием $I_{\xi i}^{(su)}$ одним из базовых семантических отношений. Для проверки этого положения подсистема (3) управления режимом самообучения и извлечения знаний формирует запрос на поиск предложения в базе (8) стохастически индексированных лингвистических текстов, в которое входят указанные индексы, связанные отношением **тема → рема:**

$$I_{\xi k}^{(su)} \rightarrow I_{\xi i}^{(su)} \qquad (16а)$$

Найденное предложение поступает в интерпретатор (4) стохастически индексированного текста и правил продукций, где проверяют, соответствует ли отношение (16а) родо-видовым, агрегатным или причинно-следственным отношениям.

При невыполнения условий (16) или (16а) считается, что данный фрагмент текста нельзя использовать для формирования ответа.

Если эти условия выполняются, то переходят к проверке возможности сформировать на основе выделенных абзацев единую семантическую структуру. С этой целью, используя таблицу индексов каждого текста, содержащего предварительно выбранные абзацы, сначала формируют списки индексов словосочетаний. Данные индексы словосчетаний входят в абзац, обозначенный соответствующим индексом:

$$(I_{\xi i}^{(t)}, I_{\xi j}^{(a)}) \longrightarrow (I_{\xi i}^{(su)}), \ldots, (I_{\xi k}^{(su)}). \qquad (17)$$

Затем определяют, с какими из абзацев связан каждый данный абзац посредством идентичных индексов словосочетаний в списках указанных абзацев. На основе указанных списков для каждого индекса абзаца составляются новые списки, каждый из которых содержит индексы других абзацев, связанных с данным абзацем идентичными индексами словосочетаний. Если при этом каждый из списков содержит не менее одного индекса абзаца, входящего не менее, чем в один из других списков, то, используя прямые или транзитивные связи между списками, образуют единый список, в который входят индексы всех абзацев. В этом случае полагают, что предварительно выбранные абзацы образуют логически связанную совокупность абзацев в виде единого фрагмента текста. В противоположном случае считается, что данная совокупность абзацев не образует логическую структуру, необходимую для формирования единого фрагмента текста. При этом она исключается из процесса обработки, и переходят к предварительному выбору новых фрагментов текстов.

После определения того, что предварительно выбранные абзацы образуют единую структуру логически связанных абзацев, на основе соответствующих таблиц индексов каждого текста формируют единую таблицу текста. При этом указанные абзацы располагаются в последовательности, определяемой порядком следования входящих в них словосочетаний запроса в вопросительном предложении запроса. Полученный в результате фрагмент текста поступает в дальнейшую обработку для определения с помощью логического вывода вида семантических связей между предложениями абзацев, содержащих все словосочетания $\{I_{\xi i}^{(u)}\}$ запроса. Цель реализации указанных функций – попытка сформировать на основе полученного фрагмента текста в соответствии с описанным выше алгоритмом стохастически индексированную семантическую структуру, включающую все словосочетания запроса. Затем полученная

семантическая структура с использованием эквивалентных преобразований и логического вывода на транзитивных зависимостях в соответствии с описанным выше алгоритмом применяется для формирования семантической структуры (15) предложения, содержащего краткий ответ, релевантный запросу пользователя. При этом корректность краткого ответа обеспечивают путем формирования описанным выше порядком нескольких идентичных стохастически индексированных семантических структур (15) на основе различных, предварительно выбранных стохастически индексированных фрагментов текстовых документов.

Полученный краткий ответ вместе с вопросительным словосочетанием при этом записывается в базу знаний (9) "запрос – ответ", которая используется для обработки повторяющихся типовых запросов пользователей, а также, как описано выше, при семантическом анализе индексируемых текстов.

Если после образования семантической структуры выяснится, что между словосочетаниями $\{I_{\xi i}^{(su)}\}$ запроса в данном фрагменте текста не поддерживаются требуемые базовые семантические связи, то переходят к поиску новых текстов для формирования ответа пользователя.

В случае положительного результата логического вывода будет сформировано предложение, содержащее краткий ответ, релевантный запросу, для выдачи его пользователю в текстовом виде на заданном языке. Если при этом пользователь потребует дать ему более полный ответ, то переходят к формированию полного ответа на основе преобразования полученного ранее фрагмента текста в соответствии с описанным ниже алгоритмом.

Рассмотрим на примере порядок реализации описанного выше алгоритма формирования краткого ответа. Допустим, что после эквивалентных преобразований поступившего запроса пользователя он принял в текстовом выражении следующий вид: ***«Какая программа используется при некорректном завершении работы с компьютером в результате пропадания напряжения в сети?»*** Это обеспечило возможность предварительного выбора следующих двух логически связанных абзацев из разных текстовых документов, содержащих в совокупности все словосочетания преобразованного запроса. Первый абзац:

«На жестком диске могут возникать логические ошибки. Логические ошибки – это нарушения в файловой структуре. Для выявления логических ошибок

используется программа «Проверка диска». Логические ошибки возникают при некорректном завершении работы с компьютером».

Второй абзац: ***«В результате пропадания напряжения в сети на жестком диске возникают нарушения в файловой структуре. В этом случае используется программа «Проверка диска».***

В стохастически индексированном виде, в котором происходит реальный процесс обработки запроса и формирования краткого ответа, текст запроса имеет следующий вид:

$$I_{\xi 0}^{(p)}: I_{\xi 01}^{(su)} \wedge I_{\xi 02}^{(su)} \rightarrow I_{\xi 03}^{(su)} \rightarrow I_{\xi 04}^{(su)} \wedge I_{\xi 05}^{(su)} \wedge I_{\xi 041}^{(su)} \wedge I_{\xi 051}^{(su)} . \quad (18)$$

При этом стохастическим индексам $I_{\xi 0\, j}^{(su)}$ соответствуют следующие словосочетания:

$I_{\xi 01}^{(su)}$: = (какая программа),

$I_{\xi 02}^{(su)}$:= (программа),

$I_{\xi 03}^{(su)}$: = (используется),

$I_{\xi 04}^{(su)}$: = (при некорректном завершении),

$I_{\xi 05}^{(su))}$: = (работы с компьютером),

$I_{\xi 041}^{(su))}$: = (в результате пропадания),

$I_{\xi 051}^{(su))}$: = (напряжения в сети).

Предложения первого абзаца в стохастически индексированном виде будут представлены следующим образом:

$$I_{\xi 1}^{(p)}: I_{\xi 12}^{(su)} \rightarrow I_{\xi 13}^{(su)} \rightarrow I_{\xi 14}^{(su)}$$

$$I_{\xi 2}^{(p)}: I_{\xi 22}^{(su)} \rightarrow I_{\xi 23}^{(su)} \rightarrow I_{\xi 24}^{(su)} \quad (19)$$

$$I_{\xi 3}^{(p)}: I_{\xi 32}^{(su)} \rightarrow I_{\xi 33}^{(su)} \rightarrow I_{\xi 34}^{(su)} \wedge I_{\xi 35}^{(su)}$$

$$I_{\xi 4}^{(p)}: I_{\xi 42}^{(su)} \rightarrow I_{\xi 43}^{(su)} \rightarrow I_{\xi 44}^{(su)} \wedge I_{\xi 45}^{(su)}$$

При этом стохастическим индексам $I_{\xi i\ j}^{(su)}$ соответствуют следующие словосочетания:

$I_{\xi 12}^{(su)}$: = (логические ошибки),

$I_{\xi 13}^{(su)}$: = (могут возникать),

$I_{\xi 14}^{(su)}$: = (на жестком диске),

$I_{\xi 22}^{(su))}$: = (логические ошибки)

$I_{\xi 23}^{(su))}$: = (- это),

$I_{\xi 24}^{(su))}$: =(нарушения в файловой структуре),

$I_{\xi 32}^{(su))}$:= (программа «Проверка диска»),

$I_{\xi 33}^{(su)}$: = (используется),

$I_{\xi 34}^{(su)}$: = (для выявления),

$I_{\xi 35}^{(su)}$: = (логических ошибок),

$I_{\xi 42}^{(su)}$: = (логические ошибки),

$I_{\xi 43}^{(su))}$: = (возникают),

$I_{\xi 44}^{(su)}$: = (при некорректном завершении),

$I_{\xi 45}^{(su)}$: = (работы с компьютером).

Предложения второго абзаца в стохастически индексированном виде будут иметь следующий вид:

$$I_{\xi 5}^{(p)} : I_{\xi 52}^{(su)} \rightarrow I_{\xi 53}^{(su)} \rightarrow I_{\xi 54}^{(su)} \wedge I_{\xi 55}^{(su)} \wedge I_{\xi 551}^{(su)} \quad (20)$$

$$I_{\xi 6}^{(p)} : I_{\xi 62}^{(su)} \rightarrow I_{\xi 63}^{(su)} \rightarrow I_{\xi 64}^{(su)}$$

При этом стохастическим индексам $I_{\xi i\,j}^{(su)}$ соответствуют следующие словосочетания:

$I_{\xi 52}^{(su)}$: = (нарушения файловой структуры),

$I_{\xi 53}^{(su)}$: = (возникают),

$I_{\xi 54}^{(su)}$: = (на жестком диске),

$I_{\xi 55}^{(su)}$: = (в результате пропадания),

$I_{\xi 551}^{(su)}$: = (напряжения в сети),

$I_{\xi 62}^{(su)}$: =(программа «Проверка диска»),

$I_{\xi 63}^{(su)}$: = (используется),

$I_{\xi 64}^{(su)}$: = (для выявления),

$I_{\xi 65}^{(su)}$: = (логических ошибок),

$I_{\xi 651}^{(su)}$: = (в этом случае).

На основе приведенных выше стохастически индексированных семантических структур описанным выше порядком будет образована стохастически индексированная семантическая структура, включающая все $I_{\xi ij}^{(su)}$ словосочетаний запроса. В качестве основы выбрана структура $I_{\xi 3}^{(p)}$, которая включает группу слов ответа $I_{\xi 32}^{(su)}$, соответствующую вопросительному словосочетанию $I_{\xi 31}^{(su)}$. При этом учитывается идентичность (с точностью до основ слов) следующих индексов словосочетаний:

$$I_{\xi 02}^{(su)} = I_{\xi 31}^{(su} = I_{\xi 64}^{(su}$$

$$I_{\xi 03}^{(su)} = I_{\xi 33}^{(su)}$$

$$I_{\xi 04}^{(su)} = I_{\xi 44}^{(su)} \qquad (21)$$

$$I_{\xi 05}^{(su)} = I_{\xi 45}^{(su)}$$

$$I_{\xi 041}^{(su)} = I_{\xi 55}^{(su)}$$

$$I_{\xi 051}^{(su)} = I_{\xi 551}^{(su)}$$

$$I_{\xi 12}^{(su)} = I_{\xi 22}^{(su)} = I_{\xi 42}^{(su)}$$

$$I_{\xi 24}^{(su)} = I_{\xi 52}^{(su)}$$

В результате указанная стохастически индексированная структура будет иметь следующий вид:

$$I_{\xi 0}^{()p)} : I_{\xi 32}^{(su)} \rightarrow I_{\xi 33}^{(su)} \rightarrow I_{\xi 34}^{(su)} \wedge I_{\xi 35}^{(su)} \rightarrow I_{\xi 24}^{(su)} \rightarrow \xi_{55}^{(su)} \wedge I_{\xi 551}^{(su)} \quad (22)$$

$$\downarrow$$

$$I_{\xi 35}^{(su)} \rightarrow I_{\xi 44}^{(su)} \wedge \xi_{45}^{(su)}$$

Учитывая, отмеченную выше идентичность соответствующих индексов и тот факт, что зависимости между индексами в данной семантической структуре имеют родо-видовой и причинно-следственный характер, получим с использованием логического вывода на транзитивных зависимостях следующую структуру:

$$I_{\xi 0}^{(p)} : I_{\xi 32}^{(su)} \wedge I_{\xi 02}^{(su)} \rightarrow I_{\xi 03}^{(su)} \rightarrow I_{\xi 04}^{(su)} \wedge I_{\xi 05}^{(su)} \wedge I_{\xi 041}^{(su)} \wedge I_{\xi 051}^{(su)} \quad (23)$$

В результате будет сформирована стохастически индексированная семантическая структура краткого ответа, которая в текстовом представлении будет иметь следующее вид: ***«Программа «Проверка диска» используется при некорректном завершении работы с компьютером в результате пропадания напряжения в сети».***

Полученный краткий ответ после замены группы слов ответа ***«Программа «Проверка диска»*** на соответствующее вопросительное словосочетание ***«Какая программа»*** будет идентичен запросу:***«Какая программа используется при некорректном завершении работы с компьютером в результате пропадания***

напряжения в сети?». Это является критерием релевантности полученного краткого ответа запросу. Поэтому полученный краткий ответ может быть выдан пользователю.

Для формирования полного ответа на основе предварительно выбранного абзаца или полученного фрагмента текста отбирают только те предложения, которые были задействованы в логическом выводе при формировании краткого ответа-предложения. При этом из предложений указанных абзацев или фрагментов текстов выстраивают последовательности, обусловленные логическими связями. Порядок логических связей такой же, как при определении семантической связанности между словосочетаниями запроса. Эти словосочетания, входящие в состав разных предложений, связаны с теми словосочетаниями запроса, которые имеются в составе предложения, содержащего группу слов ответа и главное слово вопросительного словосочетания. Порядок следования цепочек предложения определяется порядком следования соответствующих им словосочетаний запроса в сформированном ранее кратком предложении – ответе пользователю. В процессе формирования полного ответа для обеспечения согласования предложений могут производить эквивалентные преобразования отдельных предложений путем замены частей речи или членов предложений без изменения смыслового содержания этих предложений. Если эквивалентные преобразования предложения требуют замены предлогов, то их производят с учетом того, какие характеристики должны иметь части речи при сочетании их с конкретными предлогами. В случае необходимости для согласования существительных или прилагательных, местоимений или причастий с новыми предлогами могут производить замену падежей указанных частей речи. Для этого используют соответствующие правила, связывающие предлог с падежами, в которых указанные части речи согласуются с данным предлогом.

Если вопросительное слово или словосочетание запроса (как? каким образом?) предполагает не короткий ответ в одном предложении, а представление последовательности действий или описаний какого-либо процесса или явления, в этом случае короткий ответ может быть предложением-зачином, содержащим группу слов ответа типа: «следующим образом», «таким образом». При этом в следующих предложениях ответа раскрывается содержание последовательности действий или описаний, содержащих ответ пользователю с требуемой полнотой. В случае отсутствия такой типовой группы слов ответа она может быть введена дополнительно для формирования предложения-зачина. После этого группа слов ответа в предложении-зачине принимается в качестве начальной темы будущего полного ответа. Далее с

помощью логического вывода выбирается последовательность предложений одного или нескольких абзацев, которые образуют совокупность семантически связанных предложений полного ответа на данный вопрос пользователя. При этом границы ответа будут определяться непрерывной цепочкой логически связанных предложений, которая завершается при окончании одного из абзацев, если тема последнего предложения этого абзаца не связана с ремой первого предложения последующего абзаца. После формирования фрагмента текста, содержащего полный ответ, включая предложение-зачин, он выдается пользователю.

Разработанный способ может быть использован для синтеза самообучающейся системы извлечения знаний из текстовых документов поисковых систем на заданном иностранном языке. Автоматическое обучение системы правилам морфологического, синтаксического и семантического анализа производят описанным выше порядком с использованием стохастически индексированных лингвистических текстов на заданном иностранном языке. Полученные правила, также представленные на заданном иностранном языке, стохастически индексируют и записывают в соответствующие базы знаний (12-14) морфологического, синтаксического и семантического анализа. При этом производят заполнение базы данных (7) стохастически индексированных словарей базового и новых слов, а также баз (10) стохастически индексированных текстовых документов по заданным темам на данном иностранном языке.

После заполнения указанных баз данных и знаний описанным выше порядком осуществляют преобразования запросов пользователей на данном иностранном языке, предварительный выбор фрагментов текстовых документов по соответствующим темам. Затем осуществляют эквивалентные преобразования данных фрагментов текстовых документов, образование стохастически индексированных семантических структур и логический вывод с использованием указанных структур для формирования краткого ответа, релевантного запросу на заданном иностранном языке.

Разработанный способ может быть использован также для синтеза самообучающейся системы извлечения знаний из текстовых документов поисковых систем на любом из множества заданных иностранных языков. Для этой цели используют описанный выше механизм самообучения в виде стохастически индексированной системы искусственного интеллекта, основанной на применении уникальных комбинаций двоичных сигналов стохастических индексов информации для

стохастической индексации и поиска фрагментов лингвистических текстов на заданном базовом языке, содержащих описание процедур грамматического и семантического анализа. Данный механизм обеспечивает автоматическое обучение системы правилам грамматического и семантического анализа путем эквивалентных преобразований стохастически индексированных фрагментов текста на любом из заданных иностранных языков, логического вывода и формирования из указанных фрагментов текста связанных семантических структур, их стохастического индексирования для представления в формате правил продукций.

Сначала с помощью описанного выше механизма производят морфологический анализ и стохастическое индексирование лингвистических текстов на заданном базовом языке в электронном виде с одновременным автоматическим обучением системы правилам морфологического анализа. Это осуществляется одновременно с формированием базы данных (7) стохастически индексированных словарей и формированием таблиц индексов лингвистических текстов базы (8) для каждого из заданных иностранных языков, а также базы знаний (12) морфологического анализа, содержащей полученные правила продукций для заданного базового языка и каждого из заданных иностранных языков.

После этого производят морфологический и синтаксический анализ, а также стохастическое индексирование текстовых документов по заданной теме на каждом из заданных иностранных языков после получения их в электронном виде из поисковой системы. При этом производят формирование таблиц индексов текстовых документов по заданной теме и запись их в базу (10) стохастически индексированных текстов с одновременным автоматическим обучением системы правилам синтаксического анализа. Указанное обучение производят описанном выше порядком с использованием стохастически индексированных лингвистических текстов на заданном базовом языке. При этом осуществляют формирование базы знаний (13) синтаксического анализа для базового языка и каждого из заданных иностранных языков.

Затем производят семантический анализ стохастически индексированных текстовых документов по заданной теме на заданном базовом языке в электронном виде с одновременным автоматическим обучением системы правилам семантического анализа и формированием базы знаний (14) семантического анализа для базового языка и каждого из заданных иностранных языков.

После заполнения базы знаний (11-12) система переходит из режима автоматического обучения в режим обработки запросов пользователей. При этом запрос пользователя формируют на естественном заданном иностранном языке и представляют его в электронном виде после стохастического индексирования в форме вопросительного предложения, включающего вопросительное словосочетание и словосочетания, которые определяют семантику запроса. После этого описанным выше порядком преобразуют запрос пользователя в стохастически индексированном виде во множество новых запросов, эквивалентных исходному запросу на заданном иностранном языке. Затем в соответствии с запросом пользователя осуществляют предварительный выбор стохастически индексированных фрагментов текстовых документов на заданном иностранном языке в электронном виде, содержащих в совокупности все словосочетания преобразованного запроса. Используя указанные фрагменты текстовых документов формируют стохастически индексированную семантическую структуру. На основе сформированной стохастически индексированной семантической структуры с помощью логического вывода, обеспечивающего связь стохастически индексированных элементов различных текстов, и эквивалентного преобразования текста формируют краткий ответ системы, содержащий словосочетания в стохастически индексированном виде, которые определяют семантику запроса, а также группу слов ответа, соответствующую вопросительному словосочетанию запроса. При этом обеспечивают корректность краткого ответа путем формирования нескольких идентичных стохастически индексированных семантических структур на основе различных, предварительно выбранных стохастически индексированных фрагментов текстовых документов.

Затем проверяют релевантность полученного краткого ответа системы запросу посредством замены группы слов ответа на соответствующее вопросительное словосочетание в стохастически индексированном виде, получения стохастически индексированного вопросительного предложения, сравнения полученного вопросительного предложения с запросом. На основе сравнения указанных предложений при идентичности полученного вопросительного предложения и запроса принимают решение о релевантности краткого ответа системы запросу и представляют его на заданном иностранном языке.

Рассмотрим теперь другой порядок применения данного способа для синтеза самообучающейся системы, обеспечивающей одновременное извлечение знаний из текстовых документов на любом из заданных иностранных языков. В этом случае сначала производят автоматическое обучение системы описанным выше порядком правилам морфологического, синтаксического и семантического анализа с использованием стохастически индексированных лингвистических текстов на заданном базовом языке. При этом в состав базы (8) стохастически индексированных лингвистических текстов включают учебно-методические пособия по изучению каждого из заданных иностранных языков на выбранном базовом языке. В базу (11) стохастически индексированных словарей иностранных слов записывают словари, обеспечивающие прямой и обратный перевод отдельных слов с базового языка на любой из заданных иностранных языков. Затем осуществляют формирование базы данных (7) стохастически индексированного словаря и баз знаний (12-14) морфологического, синтаксического, семантического анализа на заданном базовом языке. После этого подсистема управления (3) режимом автоматического обучения осуществляет автоматическое формирование запросов к указанным базам данных и знаний для предварительного выбора фрагментов лингвистических текстов на базовом языке, содержащих знания, необходимые для изучения каждого из заданных иностранных языков. Затем производят эквивалентные преобразования текстов, формирования стохастически индексируемых семантических структур и логический вывод на заданных структурах для формирования ответов, релевантных автоматическим запросам. Эти ответы используют для формирования правил продукций морфологического, синтаксического и семантического анализа текстовых документов для каждого иностранного языка. Например, если базовым языком является русский язык, то при формировании базы знаний синтаксического анализа для изучения английского языка, среди автоматически формируемых правил могут быть следующие:

1.Если существительное без предлога стоит в начале предложения,
и это существительное стоит перед существительным с предлогом **of (in, from),**
и за этим существительным следует глагол,
то первое существительное – подлежащее.
Например: **The work of the engineer is on the table.**

2.Если словосочетание состоит из глагола-связки (глагол **to be** в личной форме) и именной части, выраженной прилагательным,

то это словосочетание – составное именное сказуемое.

Например:**The tree is big.**

Полученные правила после стохастического индексирования записывают в базы знаний (12-14) морфологического, синтаксического и семантического анализа для обеспечения извлечения знаний из текстовых документов на заданном иностранном языке в соответствии с запросами пользователей. При этом формирование базы данных стохастически индексированных словарей и таблиц индексированных текстовых документов по заданным темам производят с использованием соответствующего иностранного языка. Отметим, что в процессе семантического анализа текстовых документов по заданным темам на соответствующем иностранном языке для определениия вида семантического отношения осуществляется перевод отдельных словосочетаний с помощью базы (11) стохастически индексированных словарей иностранных слов на базовый язык. Указанное словосочетание с помощью логического вывода по таблицам индексов толковых словарей на базовом языке соотносят с одним из видов семантических отношений, индексы которых записаны в интерпретаторе (4) стохастически индексированных текстов и правил продукций. Это позволяет использовать семантический анализ для уточнения описанным выше порядком принадлежности слов к членам предложения, а также для определения вида отношений между словосочетаниями при формировании стохастически индексированной семантической структуры ответа на запрос.

С помощью указанных баз данных и знаний по командам подсистемы (3) управления режимом самообучения и извлечения знаний осуществляют эквивалентное преобразование запросов пользователей на заданных иностранных языках. Затем производят предварительный выбор фрагментов текстовых документов по заданным темам, их эквивалентные преобразования, формирование стохастически индексированных семантических структур и логический вывод на данных структурах. Это обеспечивает формирование ответов, релевантных запросам пользователей, на каждом из числа заданных иностранных языков.

Если при обработке запроса выясняется, что необходимо обращение к поисковой системе для ввода новых текстовых документов на одном из иностранных языков по заданной теме, то подсистема (3) управления режимом самообучения и извлечения знаний подключает многоязычный лингвистический процессор (1). В него поступает

команда на ввод новых документов с указанием темы и наименования языка, которые представлены на базовом языке. Многоязычный лингвистический процессор (1) с помощью базы (11) стохастически индексированных словарей иностранных слов выбирает соответствующий словарь и производит перевод слов, обозначающих наименование темы, на соответствующий иностранный язык. По полученной информации многоязычный лингвистический процессор (1) формирует формализованный запрос на заданном языке к поисковой системе для ввода новых документов на иностранном языке по соответствующей теме. Указанные документы поступают в подсистему (2) стохастического индексирования текстовых документов и выделения фрагментов текстов для описанной выше обработки и ввода в базу (10) стохастически индексированных текстовых документов по заданным темам.

Промышленная применимость

Способ синтеза самообучающейся системы извлечения знаний из текстовых документов поисковых систем прежде всего может быть использован для создания на базе Internet глобальной индустрии знаний с использованием многоязычных систем извлечения знаний из текстов. Это обеспечит качественно новый информационный сервис в различных сферах – производственной, научной, образовательной, культурной и бытовой деятельности человека с учетом современных требований развития цивилизованного общества. Другим перспективным направлением промышленного применения указанного способа являются мобильные системы (мобильный Internet). Это обусловлено возможностью создания интеллектуальных информационно-поисковых систем, обеспечивающих извлечение из больших объемов текстовых документов Internet конкретных знаний и сведений по запросам пользователей с минимизацией времени передачи и восприятия пользователем необходимой ему информации. При этом запросы могут вводиться пользователем в систему на естественном языке и в речевой форме. Важным направлением промышленного применения предложенного способа является создание нового поколения интеллектуальных обучающих систем по различным предметам и проблемным областям.

Таблица 1. Фрейм предложения

Вопросы к простым предложениям	**Вопросы к простым предложениям формируются на основе базы знаний синтаксического анализа**
Наименование простых предложений в составе сложносочиненных или сложноподчиненных	**Характеристики простых предложений**
Вопросы к группам членов предложения	**Вопросы к группам членов предложения формируются на основе вопросов к членам предложения, являющимся основой данной группы**
Наименования групп членов предложения	**Группа подлежащего, группа сказуемого, группа дополнения, группа обстоятельства, группа обособленных членов предложения, группа вводных слов, словосочетаний и вставных конструкций**
Вопросы к членам предложения	**По формату словаря (включая предлоги) и таблице перевода вопросов к частям речи в вопросы к членам предложения**
Наименования членов предложения	**Подлежащее, сказуемое (простое глагольное, составное глагольное, составное именное), определение (согласованное, несогласованное), дополнение (прямое, косвенное), обстоятельство (образа действия, места, времени, меры или степени, причины, цели, условия, уступки)**
Вопросы к частям речи	**По формату словаря**
Части речи и их характеристики	**По формату словаря**
Слово	**В контексте предложения**
Стохастические индексы основ слов	**Вычисляются по специальному алгоритму или выделяются из формата словаря**

Таблица 2. Индексы текста

Индексы основ слов	Индексы абзацев			
	$I_{\xi 1}^{(a)}$	$I_{\xi 2}^{(a)}$	...	$I_{\xi n}^{(a)}$
$I_{\xi 1}^{(u)}$	$I_{\xi 11}^{(s)}$	$I_{\xi 12}^{(s)}$	...	$I_{\xi 1n}^{(s)}$
$I_{\xi 2}^{(u)}$	$I_{\xi 21}^{(s)}$	$I_{\xi 22}^{(s)}$	...	$I_{\xi 2n}^{(s)}$
...	...	...	...	...
$I_{\xi m}^{(u)}$	$I_{\xi m1}^{(s)}$	$I_{\xi m2}^{(s)}$	...	$I_{\xi mn}^{(s)}$

Таблица 3. Индексы текстов по данной теме

Индексы основ слов	Индексы текстов			
	$I_{\xi 1}^{(t)}$	$I_{\xi 2}^{(t)}$	...	$I_{\xi n}^{(t)}$
$I_{\xi 1}^{(u)}$	$I_{\xi 11}^{(s)}$	$I_{\xi 12}^{(s)}$	...	$I_{\xi 1n}^{(s)}$
$I_{\xi 2}^{(u)}$	$I_{\xi 21}^{(s)}$	$I_{\xi 22}^{(s)}$	...	$I_{\xi 2n}^{(s)}$
...	...	...	...	...
$I_{\xi m}^{(u)}$	$I_{\xi m1}^{(s)}$	$I_{\xi m2}^{(s)}$	...	$I_{\xi mn}^{(s)}$

Формула изобретения

1. Способ синтеза самообучающейся системы извлечения знаний на заданном языке из текстовых документов поисковых систем, при котором:

обеспечивают механизм самообучения в виде стохастически индексированной системы искусственного интеллекта, основанной на применении уникальных комбинаций двоичных сигналов стохастических индексов информации,

обеспечивают автоматическое обучение системы правилам грамматического и семантического анализа путем применения эквивалентных преобразований стохастически индексированных фрагментов текста, логического вывода и формирования из них связанных семантических структур и стохастического индексирования для представления в формате правил продукций,

производят морфологический анализ и стохастическое индексирование лингвистических текстов в электронном виде с одновременным автоматическим обучением системы правилам морфологического анализа,

производят морфологический и синтаксический анализ, а также стохастическое индексирование текстовых документов по заданной теме в электронном виде на заданном языке с одновременным автоматическим обучением системы правилам синтаксического анализа,

производят семантический анализ стохастически индексированных текстовых документов по заданной теме в электронном виде с одновременным автоматическим обучением системы правилам семантического анализа,

формируют запрос пользователя на естественном заданном языке и представляют его в электроном виде после стохастического индексирования в форме вопросительного предложения,

преобразуют запрос пользователя в стохастически индексированном виде во множество новых запросов, эквивалентных исходному запросу,

в соответствии с запросом пользователя осуществляют предварительный выбор стохастически индексированных фрагментов текстовых документов в электронном виде, содержащих в совокупности все словосочетания преобразованного запроса,

формируют стохастически индексированную семантическую структуру с использованием указанных фрагментов текстовых документов,

на основе указанной структуры с помощью логического вывода, обеспечивающего связь стохастически индексированных элементов различных текстов, и эквивалентного преобразования текста формируют краткий ответ системы,

проверяют релевантность полученного краткого ответа системы запросу путем формирования на его основе вопросительного предложения, сравнения полученного вопросительного предложения с запросом,

при идентичности полученного вопросительного предложения и запроса принимают решение о релевантности краткого ответа системы запросу и представляют его на заданном языке.

2. Способ синтеза самообучающейся системы извлечения знаний на любом из заданных иностранных языках из текстовых документов поисковых систем, при котором:

обеспечивают механизм самообучения в виде стохастически индексированной системы искусственного интеллекта, основанной на применении уникальных комбинаций двоичных сигналов стохастических индексов информации для стохастической индексации и поиска фрагментов лингвистических текстов на заданном базовом языке, содержащих описание процедур грамматического и семантического анализа, и автоматического обучения системы правилам грамматического и семантического анализа путем эквивалентных преобразований стохастически индексированных фрагментов текста, логического вывода и формирования из них связанных семантических структур, их стохастического индексирования для представления в формате правил продукций,

производят морфологический анализ и стохастическое индексирование лингвистических текстов на заданном базовом языке в электронном виде с одновременным автоматическим обучением системы правилам морфологического анализа, формированием базы данных стохастически индексированных словарей и формированием таблиц индексов лингвистических текстов для каждого из заданных иностранных языков, а также базы знаний морфологического анализа, содержащей полученные правила продукций для заданного базового языка и каждого из заданных иностранных языков,

производят морфологический и синтаксический анализ, а также стохастическое индексирование текстовых документов по заданной теме на каждом из заданных иностранных языков в электронном виде из поисковой системы с представлением их в

виде таблиц индексов текстовых документов по заданной теме и записью в базы стохастически индексированных текстов с одновременным автоматическим обучением системы правилам синтаксического анализа с использованием стохастически индексированных лингвистических текстов на заданном базовом языке и формированием базы знаний синтаксического анализа для базового языка и каждого из заданных иностранных языков,

производят семантический анализ стохастически индексированных текстовых документов по заданной теме на заданном базовом языке в электронном виде с одновременным автоматическим обучением системы правилам семантического анализа и формированием базы знаний семантического анализа для базового языка и каждого из заданных иностранных языков,

формируют запрос пользователя на естественном заданном иностранном языке и представляют его в электроном виде после стохастического индексирования в форме вопросительного предложения, включающего вопросительное словосочетание и словосочетания, которые определяют семантику запроса,

преобразуют запрос пользователя в стохастически индексированном виде во множество новых запросов, эквивалентных исходному запросу на заданном иностранном языке,

в соответствии с запросом пользователя осуществляют предварительный выбор стохастически индексированных фрагментов текстовых документов на заданном иностранном языке в электронном виде, содержащих в совокупности все словосочетания преобразованного запроса,

формируют стохастически индексированную семантическую структуру на основе указанных фрагментов текстовых документов,

на основе сформированной стохастически индексированной семантической структуры с помощью логического вывода, обеспечивающего связь стохастически индексированных элементов различных текстов, и эквивалентного преобразования текста формируют краткий ответ системы, содержащий словосочетания в стохастически индексированном виде, которые определяют семантику запроса, а также группу слов ответа, соответствующую вопросительному словосочетанию запроса,

проверяют релевантность полученного краткого ответа системы запросу путем замены группы слов ответа на соответствующее вопросительное словосочетание в стохастически индексированном виде, получения стохастически индексированного

вопросительного предложения, сравнения полученного вопросительного предложения с запросом и при идентичности полученного вопросительного предложения и запроса принимают решение о релевантности краткого ответа системы запросу и представляют его на заданном иностранном языке.

3. Способ по п.1 или 2, отличающийся тем, что при неудачной попытке сформировать вопросительное предложение, идентичное запросу пользователя, запрашивают новые текстовые документы из поисковой системы для поиска ответа, релевантного запросу пользователя,

4. Способ по любому из пунктов 1-3, отличающийся тем, что дополнительно по запросу пользователя формируют полный ответ, содержащий более подробную информацию или совокупность конкретных знаний, при этом используют логический вывод для образования стохастически индексированной семантической структуры и необходимые эквивалентные преобразования указанной совокупности фрагментов текстов для получения стохастически индексированного нового текста, раскрывающего с возможной детализацией содержание полученного ранее краткого ответа.

5. Способ по п. 1 или 2, отличающийся тем, что автоматическое обучение системы правилам морфологического анализа производят путем выделения в стохастически индексируемом тексте определенного набора словоформ каждого слова, получения стохастических индексов основы слова и заданного набора его окончаний или предлогов, произвольного доступа по указанным индексам к стохастически индексированным лингвистическим текстам, выделения из них фрагментов, связывающих указанный набор окончаний слова или предлогов с соответствующей данному слову частью речи, а также с полным набором окончаний или предлогов, получаемых при склонении или спряжении, преобразования данных фрагментов в формат правил продукций путем их стохастического индексирования, обеспечивая при этом корректность каждого правила путем независимого его формирования на основе нескольких фрагментов из соответствующих лингвистических текстов, и получения таблицы индексов правил продукций для базы знаний морфологического анализа.

6. Способ по любому из п.п.2-5, отличающийся тем, что при стохастическом индексировании лингвистических текстов после определения части речи каждого слова с помощью правил базы знаний морфологического анализа заполняют базу

данных стохастически индексированного словаря стохастическими индексами основы каждого очередного слова и полного набора его окончаний или предлогов.

7. Способ по любому из п.п. 2-6, отличающийся тем, что для формирования таблиц индексов текстов осуществляют стохастическое преобразование информации и получение уникальных двоичных комбинаций индексов основ слов, их окончаний, предлогов, предложений, абзацев и названий текстов, которые помещают в таблицы индексов базы стохастически индексированных текстов с обеспечением связности между указанными индексами, определенной в исходном тексте и обеспечивающей его восстановление по таблице индекса.

8. Способ по п.1 или 2, отличающийся тем, что автоматическое обучение системы правилам синтаксического анализа осуществляют путем поиска в стохастически индексированных лингвистических текстах фрагментов, описывающих порядок синтаксического разбора предложений, при этом реализуется логический вывод для получения стохастически индексированной семантической структуры, определяющей связь синтаксических элементов и структур с заданными частями речи слов, и формирования правил продукций, определяющих синтаксический разбор предложений по морфологическим характеристикам слов, обеспечивая при этом корректность каждого правила путем независимого его формирования на основе нескольких фрагментов из соответствующих лингвистических текстов, полученные правила заносят в базу знаний синтаксического анализа, по мере заполнения которой осуществляют ее стохастическое индексирование и представление в виде таблицы индексов.

9. Способ по п.1 или 2, отличающийся тем, что автоматическое обучение системы правилам семантического анализа текста осуществляют путем формирования запроса к таблицам индексов лингвистических текстов по стохастическим индексам основ слов и частей речи, не точно определенных членов предложения, и получения ответа в виде фрагмента текста, описывающего семантические характеристики, которыми должны обладать слова для их соответствия данному конкретному члену предложения, и по полученному ответу, используя стохастический индекс основы данного слова и требуемые семантические характеристики, обращаются к таблицам индексов толковых словарей и энциклопедий общего и тематического назначения, при этом с помощью логического вывода делают попытку образовать стохастически индексированную семантическую структуру, связывающую данное слово и требуемые

семантические характеристики, в положительном случае считают, что указанный член предложения определен точно, а фрагмент текста, релевантный запросу, преобразуют в правило продукций, обеспечивая при этом корректность каждого правила путем независимого его формирования на основе нескольких фрагментов из соответствующих лингвистических текстов, которое включают в базу знаний семантического анализа, стохастически индексируют данную базу, представляют в виде таблицы индексов и применяют при семантическом анализе слов, как членов предложения, и отношений между словами, выраженных словосочетаниями.

10. Способ по любому из п.п.2-9, отличающийся тем, что после образования таблицы индексов каждого текста и завершения его морфологического, синтаксического и семантического анализа формируют стохастические индексы наименований частей речи, членов предложения и вопросов к ним, которые соответствуют каждому слову в составе предложений, и записывают указанные индексы в ячейки таблицы индексов данного текста, что позволяет при поиске фрагментов текста автоматически определять, к какой части речи, члену предложения относится каждое слово, и формировать вопросы к нему.

11. Способ по любому из п.п.2-10, отличающийся тем, что после получения всех таблиц индексов текстов формируют таблицу индексов текстов по данной теме, строки которой поименованы неповторяющимися стохастическими индексами основ слов, а каждый столбец соответствует стохастическому индексу конкретного текста, при этом в ячейки таблицы записывают стохастические индексы абзацев, в которых в данном тексте содержится слово с соответствующим индексом основы, полученную таблицу индексов по данной теме применяют для предварительного поиска фрагментов, содержащих определенную совокупность словосочетаний запроса.

12. Способ по любому из п.п.1-11, отличающийся тем, что эквивалентные преобразования исходного запроса пользователя осуществляют с использованием синонимов, близких по смыслу слов, а также замены частей речи и членов предложения с сохранением смыслового содержания исходного запроса на основе применения стохастически индексированных правил морфологического, синтаксического и семантического анализа для получения эквивалентных структур словосочетаний вопросительного предложения запроса и сохранения семантической связи между ними.

13. Способ по любому из п.п.1-12, отличающийся тем, что совокупность семантически связанных фрагментов текста, содержащих все слова запроса

пользователя, формируют путем обращения по стохастическим индексам указанных основ слов к таблице индексов текстов по заданной теме, выбора стохастических индексов абзацев и соответствующих им текстов, содержащих в совокупности все словосочетания запроса, обращения по указанным индексам к таблице индексов каждого из выбранных текстов, логического вывода по таблицам индексов и эквивалентных преобразований текстов для образования стохастически индексированной семантической структуры, связывающей индексы группы слов ответа, соответствующего вопросительному словосочетанию запроса, а также все словосочетания запроса, определяющие семантику запроса и входящие в предварительно выбранные абзацы.

14. Способ по п.13, отличающийся тем, что успешно сформированная в процессе логического вывода стохастически индексированная семантическая структура, соответствующая запросу пользователя, принимается в качестве основы для формирования с использованием полученной совокупности фрагментов текста вопросительного предложения, идентичного запросу пользователя, которое образуют путем эквивалентного преобразования стохастических индексов основ слов запроса и их окончаний с помощью правил баз знаний для обеспечения требуемых семантических характеристик каждого словосочетания текстового фрагмента, входящего в состав запроса, а также с использованием логического вывода на транзитивных зависимостях между словосочетниями для объединения их в единое вопросительное предложение, идентичное запросу пользователя, которое содержит группу слов ответа, соответствующую вопросительному словосочетанию запроса.

15. Способ по любому из п.п. 1-14, отличающийся тем, что корректность краткого ответа обеспечивают путем формирования нескольких идентичных стохастически индексированных семантических структур упомянутого ответа на основе различных, предварительно выбранных стохастически индексированных фрагментов текстовых документов.

16. Способ по любому из п.п. 1-15, отличающийся тем, что в процессе поиска и формирования ответа с использованием таблиц индексов текстовых документов самообучение системы осуществляют путем формирования индексированных текстовых элементов, связывающих запрос и релевантный краткий ответ, для получения базы знаний, содержащей элементы типа «запрос – ответ», которую стохастически индексируют, представляют в виде таблицы индексов и применяют при

грамматическом и семантическом анализе предложений текста, а также при формировании ответов на повторяющиеся запросы пользователей, содержащиеся в указанной индексированной базе знаний.

17. Способ по любому из п.п. 4-16, отличающийся тем, что для формирования полного ответа, содержащего знания, релевантные запросу пользователя, на основе краткого ответа с помощью логического вывода по таблицам индексов, использованных при получении фрагмента текста, формируют стохастически индексированную семантическую структуру, связывающую группу слов ответа со стохастическими индексами основ слов предложений, поддерживающих транзитивную зависимость, обеспечивающую в своей совокупности полное раскрытие содержания краткого ответа в рамках сформированного фрагмента текста, затем с помощью эквивалентных преобразований предложений на основе указанной стохастически индексированной семантической структуры получают единый связанный текст полного ответа.

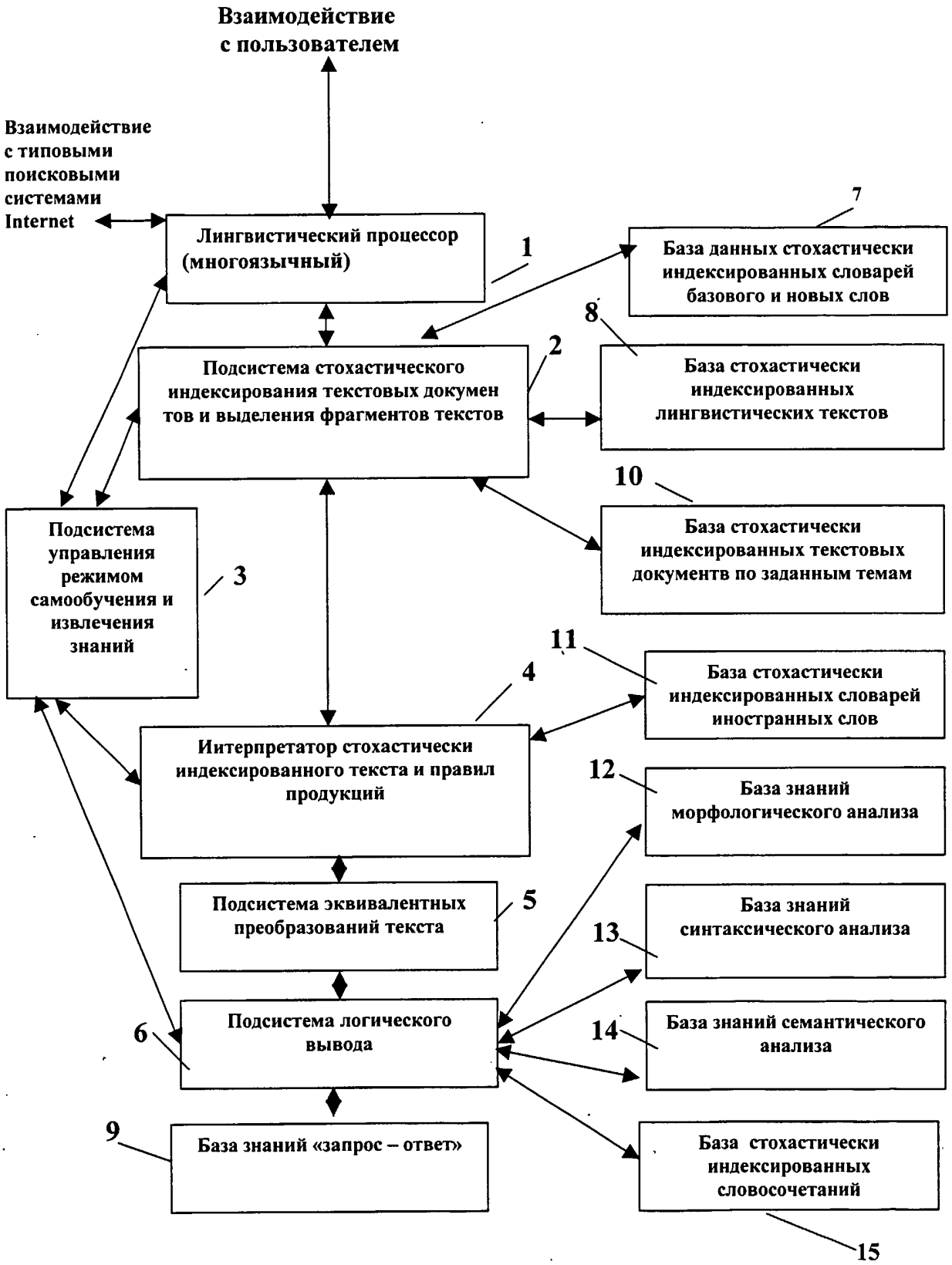
18. Способ по любому из п.п. 1-17, отличающийся тем, что эквивалентное преобразование стохастически индексированных фрагментов текста производят путем представления каждого предложения в виде совокупности стохастически индексированных словосочетаний, которые преобразуют с использования правил баз знаний морфологического, синтаксического и семантического анализа путем эквивалентного преобразования стохастических индексов основ однокоренных слов, их окончаний и предлогов для образования новых частей речи или членов предложения с обеспечением неизменности связи указанных словосочетаний в рамках стохастически индексированной семантической структуры каждого предложения и согласования указанных предложений между собой при образовании из них нового фрагмента текста.

19. Способ по любому из п.п. 1-18, отличающийся тем, что при появлении в процессе стохастического индексирования текстовых документов в индексируемом тексте нового слова, не содержащегося в словаре стохастически индексированных слов и в лингвистических текстах, находят в данном словаре однокоренное слово с указанным новым словом, а в базе знаний морфологического анализа находят правила для эквивалентного преобразования найденного в словаре однокоренного слова в новое слово, при этом по виду эквивалентного преобразования определяют часть речи, к которой относится новое слово и все его словоформы, получаемые при

склонении или спряжении, а при отсутствии однокоренных слов в словаре выбирают из текста определенный набор словоформ нового слова, по предлогам или окончаниям которых с помощью стохастически индексированного словаря или правил продукций морфологического анализа определяют часть речи, к которой оно относится, и полный набор его словоформ, получаемых при склонении или спряжении.

20. Способ по любому из п.п. 2-19, отличающийся тем, что для одновременного извлечения знаний из текстовых документов на заданных иностранных языках сначала осуществляют автоматическое обучение системы правилам морфологического, синтаксического, семантического анализа для заданного базового языка, производят формирование базы стохастически индексированного словаря и баз знаний морфологического, синтаксического, семантического анализа с использованием стохастически индексированных лингвистических текстов на заданном базовом языке, с помощью сформированных баз осуществляют автоматическое формирование запросов для автоматического обучения системы любому из заданных иностранных языков, при этом производят предварительный выбор по автоматически сформированным запросам фрагментов лингвистических текстов на базовом языке, содержащих знания, необходимые для изучения заданного иностранного языка, эквивалентные преобразования указанных текстов, формирование стохастически индексируемых семантических структур и логический вывод на заданных структурах для формирования ответов, релевантных автоматическим запросам, которые используют для формирования баз знаний морфологического, синтаксического и семантического анализа для любого из заданных иностранных языков, обеспечивающих извлечение знаний из текстовых документов на заданном иностранном языке.

1/1


Взаимодействие с пользователем
Взаимодействие с типовыми поисковыми системами Internet
Лингвистический процессор (многоязычный)
1
7
База данных стохастически индексированных словарей базового и новых слов
8
2
Подсистема стохастического индексирования текстовых документов и выделения фрагментов текстов
База стохастически индексированных лингвистических текстов
10
Подсистема управления режимом самообучения и извлечения знаний
3
База стохастически индексированных текстовых документв по заданным темам
11
4
База стохастически индексированных словарей иностранных слов
Интерпретатор стохастически индексированного текста и правил продукций
12
База знаний морфологического анализа
Подсистема эквивалентных преобразований текста
5
13
База знаний синтаксического анализа
6
Подсистема логического вывода
14
База знаний семантического анализа
9
База знаний «запрос – ответ»
База стохастически индексированных словосочетаний
15

ФИГ. 1

INTERNATIONAL SEARCH REPORT

International application No.
PCT/RU 02/00258

A. CLASSIFICATION OF SUBJECT MATTER 7 G06F 17/30, G09B 19/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols) 7

G06F 17/00, 17/20, 17/21, 17/27, 17/28, 17/30, 17/40, G09B 19/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	NASIPNIY V.V. et al. Postroenie intellektualnoi informatsionnoi Poiskovoi systemi. Moskva, Prometei, 2001, p. 3-24	1-20
A	NASIPNIY V.V. et al. Rasvitie teorii postroenia otkrytikh system na osnove informatsionnoi tekhnologii iskusstvennovo intellekta. Moskva, voennoe isdatelstvo, 1994, p. 36-112	1-20
A	(NAUCHNO-ISSLEDOVATELSKAYA I PROIZVODSTVENNAYA FIRMA "TEKHINTELL") 27.04.2001	1-20
A	US 5787234 A (BRUCE G. MOLLOY) Jul. 28, 1998	1-20
A	US 5454106 A (INTERNATIONAL BUSINESS MACHINES CORPORATION) Sep. 26, 1995	1-20

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
03 February 2003 (03.02.03)	**20 February 2003 (20.02.03)**
Name and mailing address of the ISA/**RU**	Authorized officer
Facsimile No.	Telephone No.

Form PCT/ISA/210 (second sheet) (July 1992)

ОТЧЕТ О МЕЖДУНАРОДНОМ ПОИСКЕ

Международная заявка № PCT/RU 02/00258

A. КЛАССИФИКАЦИЯ ПРЕДМЕТА ИЗОБРЕТЕНИЯ: G06F 17/30, G09B 19/00

Согласно международной патентной классификации (МПК-7)

B. ОБЛАСТИ ПОИСКА:

Проверенный минимум документации (система классификации и индексы) МПК-7:
G06F 17/00, 17/20, 17/21, 17/27, 17/28, 17/30, 17/40, G09B 19/00

Другая проверенная документация в той мере, в какой она включена в поисковые подборки:

Электронная база данных, использовавшаяся при поиске (название базы и, если, возможно, поисковые термины):

C. ДОКУМЕНТЫ, СЧИТАЮЩИЕСЯ РЕЛЕВАНТНЫМИ:

Категория*	Ссылки на документы с указанием, где это возможно, релевантных частей	Относится к пункту №
A	НАСЫПНЫЙ В.В. и др. Построение интеллектуальной информационной поисковой системы. Москва, Прометей, 2001, стр. 3-24	1-20
A	НАСЫПНЫЙ В.В. Развитие теории построения открытых систем на основе информационной технологии искусственного интеллекта. Москва, Военное издательство, 1994, стр. 36-112	1-20
A	RU 2166208 C2 (НАУЧНО-ИССЛЕДОВАТЕЛЬСКАЯ И ПРОИЗВОДСТВЕННАЯ ФИРМА "ТЕХИНТЕЛЛ") 27.04.2001	1-20
A	US 5787234 A (BRUCE G. MOLLOY) Jul. 28, 1998	1-20
A	US 5454106 A (INTERNATIONAL BUSINESS MACHINES CORPORATION) Sep. 26, 1995	1-20

☐ последующие документы указаны в продолжении графы C. ☐ данные о патентах-аналогах указаны в приложении

* Особые категории ссылочных документов:

A документ, определяющий общий уровень техники

E более ранний документ, но опубликованный на дату международной подачи или после нее

O документ, относящийся к устному раскрытию, экспонированию и т.д.

P документ, опубликованный до даты международной подачи, но после даты испрашиваемого приоритета и т.д.

T более поздний документ, опубликованный после даты приоритета и приведенный для понимания иззобретения

X документ, имеющий наиболее близкое отношение к предмету поиска, порочащий новизну и изобретательский уровень

Y документ, порочащий изобретательский уровень в сочетании с одним или несколькими документами той же категории

& документ, являющийся патентом-аналогом

Дата действительного завершения международного поиска: 03 февраля 2003 (03.02.2003)

Дата отправки настоящего отчета о международном поиске: 20 февраля 2003 (20.02.2003)

Наименование и адрес Международного поискового органа
Федеральный институт промышленной собственности
РФ,123995, Москва, Г-59, ГСП-5, Бережковская наб., 30,1 Факс: 243-3337, телетайп: 114818 ПОДАЧА

Уполномоченное лицо:
О. Ревинский

Телефон № 240-25-91

Форма PCT/ISA/210 (второй лист)(июль 1998)