МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 4 Москва 2021

ИНФОРМАЦИОННЫЙ ПОИСК

УДК 004.891.2: [004.774.2:002]

В.А. Трусов

Концептуальный подход к поиску и семантической обработке научно-технической информации в распределенных системах Интернета

Рассматривается концептуальный подход поиска и семантической обработки, структурированной и неструктурированной научно-технической информации в распределённых информационных системах глобальной вычислительной сети Интернет. Представлена процедурная модель организации многоитерационного человеко-машинного взаимодействия с жестко закрепленными обратными связями для сокращения совокупного времени и повышения полноты поиска и семантической обработки информации. Приведены функционально-практические основы разработки экспертно-аналитической системы поддержки процесса информационной и аналитической работы.

Ключевые слова: *поиск информации, семантическая обработка информации, структурированная информация, неструктурированная информация, экспертно-аналитическая система, научно-техническая информация*

DOI: 10.36535/0548-0027-2021-04-1

ВВЕДЕНИЕ

Научно-технический прогресс и научно-технологическое развитие не остановить. Человеческая мысль, генерируемая в процессе создания новых объектов техники и технологий, интерпретируется в виде различного рода информации (патентной, научно-технической, маркетинговой, деловой и прочей). На современном этапе формализация и распространение сущности интеллекта осуществляется принципиально иначе, и в этом кроется сакральный смысл современного понятия «информация», которое уже фактически стало сродни слову «Интернет». Великие мыслители, инженеры, математики и не предполагали, что научно-технический прогресс (НТП) дойдет до того, что любой исследователь, из любой точки мира сможет овладевать знаниями и перенимать опыт своих коллег, что неминуемо становится серьезным технологическим толчком, в том числе для ускорения НТП. Это приводит к важному выводу: «информация» и есть тот механизм, который способствовал колоссальному развитию НТП в XX и начале XXI веков, и на основе которого продолжается современное научно-технологическое развитие.

Такая информационная парадигма предопределила широкое распространение структурированных и неструктурированных источников научно-технической информации в распределённых информационных системах глобальной вычислительной сети Интернет (далее – РИС ГВС). Следует отметить, что понимание структурированной и неструктурированной информации нужно трактовать не только с точки зрения самих источников информации и ее размещения, но и с точки зрения поведения пользователя при работе в РИС ГВС с такими источниками.

Поиск и обработка научно-технической информации в РИС ГВС становится нетривиальной задачей. Для реализации информационной потребности (целей поиска информации) пользователь должен обладать достаточно серьезной квалификацией, как в сфере информационной и аналитической работы, так и во владении предметной областью, которая по сути и определяет его информационную потребность. Только такое положение дел может обеспечить повышение полноты реализации информационной потребности пользователя. Основная проблема заключается в том, что пользователь по определению не может быть квалифицированным информационным работником, так как информационная работа имеет серьезную специфику, требует определенных навыков и знаний, которыми необходимо овладевать на протяжении длительного времени. Но зато пользователь является специалистом в своей предметной области и хорошо представляет информационную потребность. Перед информационными работниками стоит противоположная задача – быстрое изучение предметной области и выявление информационной потребности, так как без понимания (смысла) того, что нужно искать, получить результат крайне сложно, а зачастую невозможно.

Ключевым показателем поиска и обработки научно-технической информации в РИС ГВС является общее время $\tilde{t}_{об}$ [1], затрачиваемое пользователем (оператором) на эти операции, что определяется распределённым характером и большим объемом хранящейся в РИС ГВС информации. Для принятия управленческих решений, а именно для снятия внешней информационной неопределенности, необходима дополнительная информация. Работа по поиску и обработке информации для решения поставленных задач должна, в конечном итоге, быть выполнена в минимально отведенное время. На общее время информационной работы влияют поставленные информационно-аналитические задачи, квалификация пользователя (оператора), вид информации, знание источников информации и их особенностей.

Общее (совокупное) время $\tilde{t}_{об}$, затрачиваемое пользователем на поиск и семантическую обработку информации в РИС ГВС, определяется по формуле:

$$\tilde{t}_{об} = \tilde{t}^{\,a}_{об} + \tilde{t}^{\,м}_{об},$$

где, $\tilde{t}^{\,a}_{об}$ – аналитическая (семантическая) составляющая, охватывает работы по анализу предметной области, лежит в основе формирования тезауруса, конструкций поисковых предписаний и отвечает за пертинентность поисковых запросов; $\tilde{t}^{\,м}_{об}$ – машинная составляющая, предназначена для поисковых работ в РИС ГВС с использованием существующих поисковых машин (*Yandex*, *Google*, *Yahoo*, *Bing*, ИРБИС и пр.) и отвечает за релевантность поисковых запросов.

Аналитическая (семантическая) составляющая, в общем понимание, есть ничто иное как просмотр вручную всего массива релевантной информации и отбор из этого массива только той информации, которая соответствует информационной потребности. Аналитическая составляющая является самой трудоемкой и напрямую влияет на общее время поиска и обработки информации в РИС ГВС. Важнейшая задача при поиске и семантической обработке информации в РИС ГВС – это снижение аналитической нагрузки на пользователя и увеличение роли машинной составляющей. Это возможно только при грамотном, системном подходе, где весь процесс поиска и семантической обработки информации представлен в виде четко формализованных стадий и этапов, понятных входных поисковых параметров и желаемых результатов. В противном случае такие работы могут проводиться бесконечно долго и не приведут к ожидаемым результатам.

ОРГАНИЗАЦИЯ ПОИСКА И СЕМАНТИЧЕСКОЙ ОБРАБОТКИ ИНФОРМАЦИИ В РИС ГВС

Решение приведенных проблем и задач предполагает следующую парадигму (системность), основанную на трехстадийном (*предпоисковом, поисковом, послепоисковом*) подходе к поиску и семантической обработке информации в РИС ГВС, который заключается в применении методов автоматизации (что необязательно, процесс может происходить в ручном режиме самостоятельно оператором) в 1-й и 3-й стадии, а на 2-й стадии используются существующие методы автоматизации (поисковые механизмы РИС ГВС и структурированных источников информации).

Следует отметить, что такая парадигма не нова, она используется повсеместно при физической работе с внешней (дополнительной) информацией и у нас, и на западе. Но вот подходы к автоматизации этой парадигмы существенно различаются: советская школа всегда особое внимание уделяла первой и третьей стадиям, была нацелена на их автоматизацию, что сделать крайне сложно, так как природу смысла информационной потребности пользователя практически невозможно представить в виде какой-то формальной структуры, отображающей действительную картину, которую в итоге желает видеть пользователь. На Западе развивался иной подход, он заключался в автоматизации второй стадии, что сделать гораздо проще и существующие поисковые системы идут именно таким путем, считая пользователя РИС ГВС заранее квалифицированным, с точки зрения информационной работы, что на практике не соответствует действительности.

Разница в подходах проста и очевидна:

- советский подход нацелен на понимание того, что нужно искать, ставя в основу предпоисковую и послепоисковую работу, так как если мы будем знать, что искать, то даже стандартными механизмами поиска, разработанными в 1970-х гг. (в действительности ничего кардинально нового еще не придумали), получим 100% результат;
- западный подход основан на том, как нужно искать, ставя в основу поисковую работу и развитие механизмов поиска информации.

Но если мы не знаем, что нужно подавать на вход поисковым системам, не понимаем требуемый результат поиска, то как бы мы ни развивали механизмы поиска, накручивая туда семантику, «искусственный интеллект», машинное обучение, нейросетевую алгоритмизацию и т.п., все равно не сможем залезть пользователю в голову и понять, что он хочет получить в итоге, отсюда аналитическая составляющая будет очень трудоёмка.

Предлагаемый в настоящей статье концептуальный подход поиска и семантической обработки информации в РИС ГВС, позволяющий снизить трудоемкость именно аналитической составляющей.

Первая ступень – анализ информационного пространства РИС ГВС, исходя из информационной потребности пользователя. Основная цель анализа – постановка задачи на поиск информации и определение начальной информационной структуры лексических единиц, описывающих предметную область, в понимании (исходя из практического опыта) пользователя. В зависимости от типа работ формируются определённые области информационного пространства:

– если работа носит исследовательский характер, то формируется только информационное пространство предмета или объекта информационной потребности;

– если работа носит прикладной характер, то необходимо понимать связь информационных пространств предмета и объекта между собой, в этом и будет заключатся информационная потребность.

Для лучшего понимания информационного пространства можно рекомендовать формировать его в виде объектно-графической модели, чтобы формализовать лексическую природу (природу смысла) информационной потребности и наглядно показать распределения групп лексических единиц и их взаимосвязей.

Вторая ступень – поиск структурированной информации (рис. 1) необходим для формирования лексического понимания и определения терминологической структуры объекта исследования. Фактически для того, чтобы получить информационный результат, нужно сначала понять, как другие внешние субъекты лексически (терминологически) определяют информационное пространства объекта и предмета информационной потребности, так как собственное лексическое представление информационной потребности может быть очень ограниченным.

В контексте настоящего концептуального подхода структурированный поиск проводится первым в силу того, что в РИС ГВС располагается большое количество известных электронных источников структурированной научно-технической информации (*ELibrary*, *Elsevier*, *Springer*, ФИПС, *USPTO*, ГПНТБ, *Espacenet*, *STN International*, и многие другие), позволяющих осуществлять быстрый доступ к информации.

Немаловажный фактор – это сама сущность информации в данных источниках, выстроенная в общепринятую архитектуру. В зависимости от вида информации архитектура включает в себя стандартный набор информационных параметров – ключевые слова, дескрипторы, авторы, держатели (патентообладатели), классификационные индексы и т.п., анализ этой информационной основы позволяет понять лексическое распределение (представление) информационной потребности внешними субъектами, ведь искать приходится именно информацию, генерируемую внешними субъектами.

Поиск проводится в базовых группах лексических единиц с применением многоитерационного подхода для общеизвестных источников структурированной информации (быстрый поиск). Средствами синтаксического анализа изучаются найденные документы, определяются в них лексические единицы, соответствующие объекту исследования, и добавляются в тезаурус пользователя. Методика учитывает формальную информационную структуру научно-технических и патентных документов, разбивая их на группы и элементы.

Состав лексических единиц (дескрипторы, ключевые слова, авторы, держатели (патентообладатели) информации, индексы систем классификации информации и т.п.) обусловлен инструментами поиска структурированной информации в РИС ГВС, не позволяющими в полной мере проводить дескрипторный поиск информации.

Основные результаты второй ступени поиска информации:

- действительное понимание лексического определения объекта исследований за счет анализа внешнего информационного пространства предметной области;
- формулирование эталонных дескрипторов для проведения неструктурированного поиска информации, так как неструктурированный поиск основывается на дескрипторном принципе.

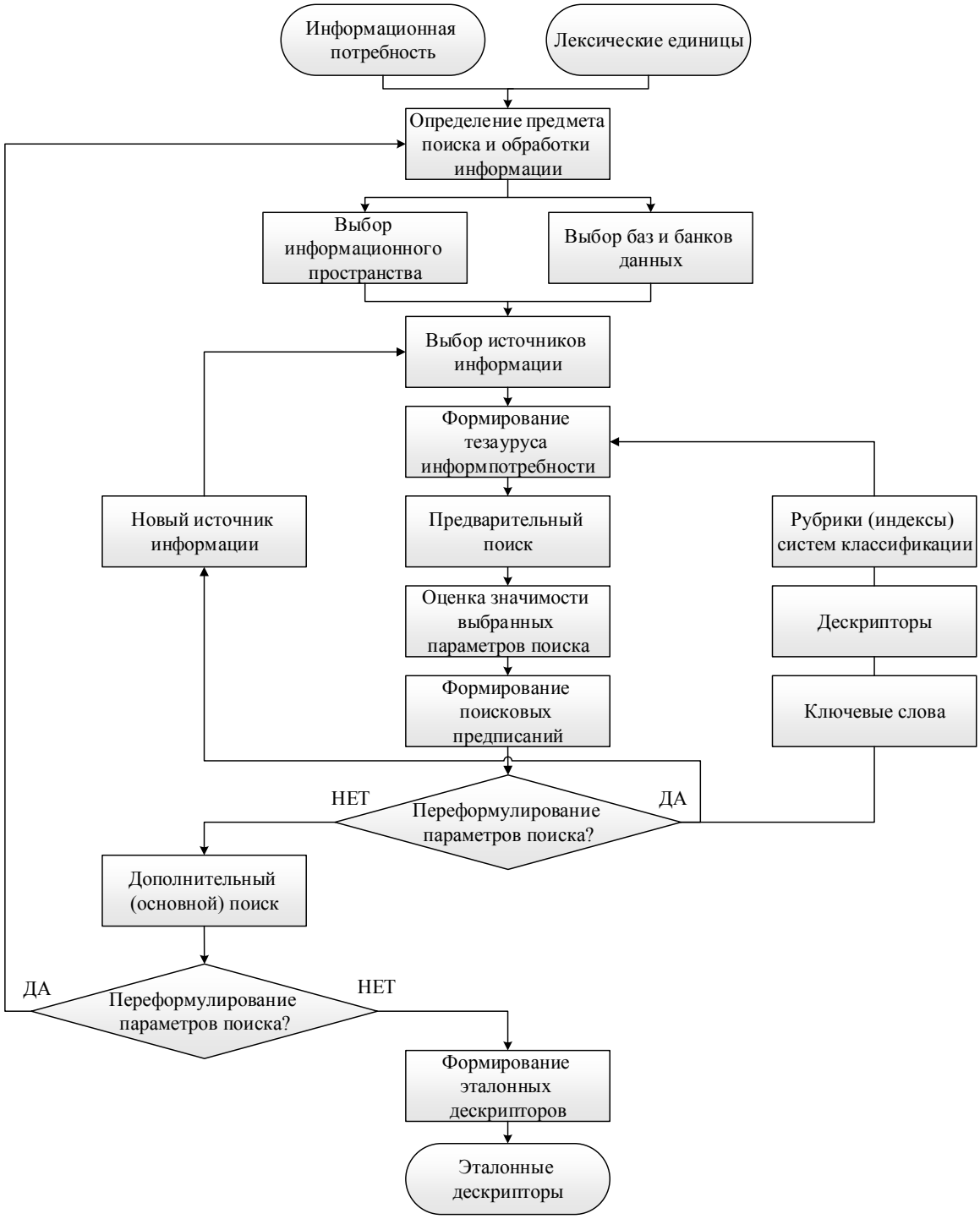


Рис. 1. Процедурная модель структурированного поиска информации в распределённых информационных системах глобальной вычислительной сети

Следует отметить, что в рамках структурированного поиска информации определяются границы поиска и обработки информации. Найденные документы, соответствующие информационной потребности пользователя, формируются в виде дайджеста, проранжированного в порядке убывания степени смыслового отношения материала к информационной потребности.

Третья ступень – поиск неструктурированной информации (рис. 2) проводится на расширенном составе всех доступных групп лексических единиц с применением многоитерационного подхода на основе существующих поисковых машин (*Google*, *Yandex*, *Yahoo*, и прочих). Использование поисковых машин РИС ГВС предопределено тем, что эти инструменты осуществляют исчерпывающее индексирование информационного пространства РИС ГВС, и разрабатывать специализированные инструменты поиска научно-технической информации нет смысла, необходимо просто правильно (грамотно) использовать существующие.

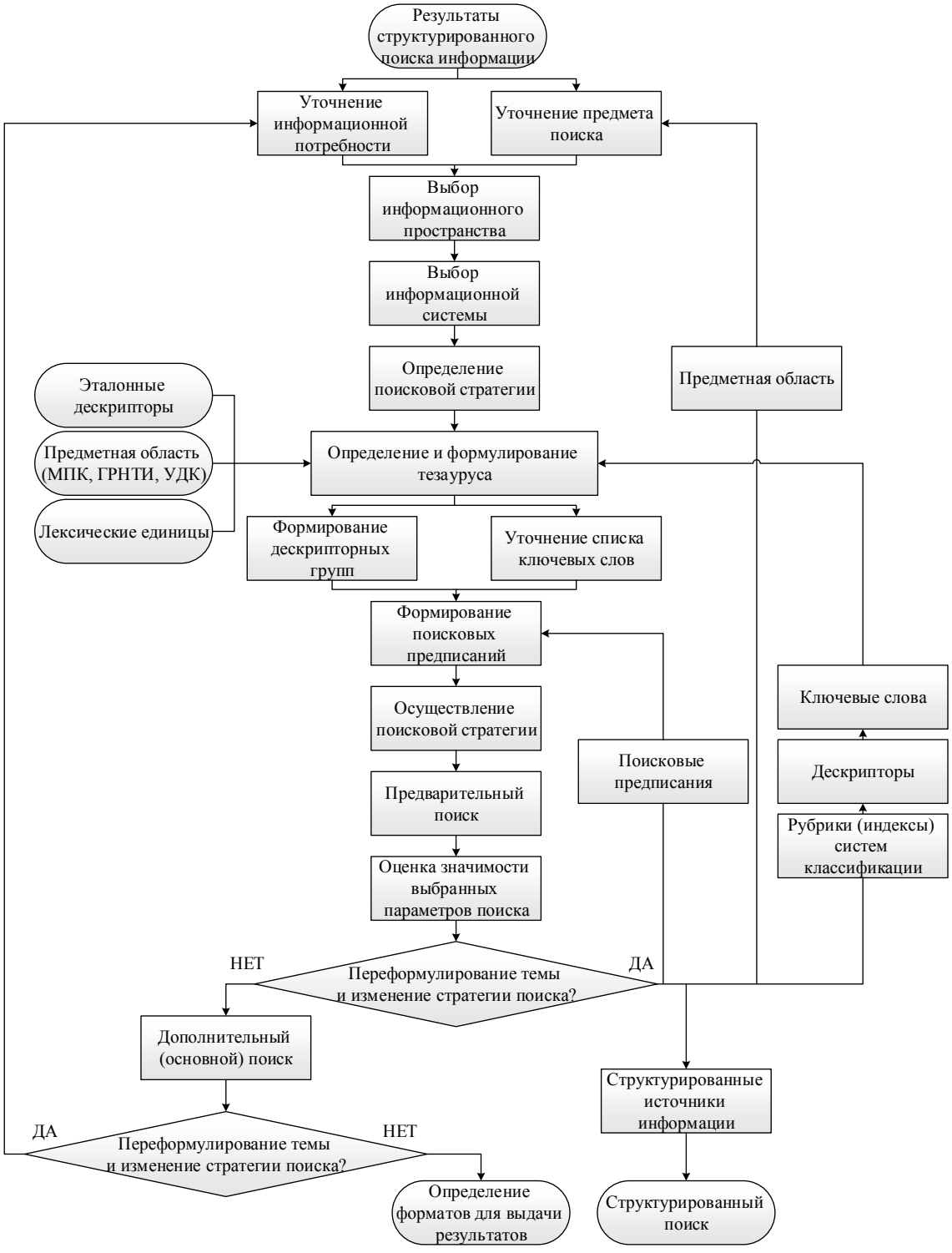


Рис. 2. Процедурная модель проведения структурированного поиска информации в распределённых информационных системах глобальной вычислительной сети

В основе дескрипторного принципа поиск неструктурированной информации лежат эталонные дескрипторы, полученные на этапе проведения структурированного поиска и определяющие границы информационной потребности. Используются следующие группы лексических единиц: дескрипторы и ключевые слова, описывающие предметную область (на основе существующих систем классификации); дескрипторы и ключевые слова, расширяющие предметную область (на основе существующих и специализированных систем классификации); дескрипторы и ключевые слова, суживающие предметную область (на основе существующих и специализированных систем классификации); ключевые слова; ключевые слова и словоформы; ключевые слова и синонимы; ключевые слова и антонимы; дескрипторы; дескрипторы сино-

нимии; дескрипторы антиномии; дескрипторы антонимии; дескрипторы с учетом расширенной предметной области; дескрипторы с учетом сужения предметной области; авторы; фирмы разработчики (патентообладатели); держатели/издатели информации и др. Такой расширенный состав групп лексических единиц обусловлен методами и подходами поиска информации, заложенными в поисковые машины РИС ГВС и позволяющими обрабатывать сложнокомбинированные поисковые запросы.

В ходе поиска определяются и добавляются в базу данных (перечень) пользователя новые источники структурированной информации, по которым в дальнейшем осуществляется поиск. Найденные документы, соответствующие информационной потребности пользователя, формируются в виде проранжированного дайджеста в порядке убывания степени отношения материала к информационной потребности. По итогам структурированного поиска информации повторяющиеся документы исключаются.

Такая структура, необходимая для более полного и корректного поиска и обработки информации в зависимости от вида её размещения, позволяет на ранних этапах выявлять лексическое понимание объекта исследования в рамках информационного пространства предметной области, существенно снижая время поиска.

ПРОЦЕДУРА ПОИСКА И СЕМАНТИЧЕСКОЙ ОБРАБОТКИ ИНФОРМАЦИИ В РИС ГВС

Последовательный многоитерационный процесс человеко-машинного (кибернетического) взаимодействия с жестко закрепленными обратными связями для повышения полноты поиска и качества обработки информации проводится идентично для структурированного и неструктурированного поиска информации и включает следующие этапы:

предпоисковая стратегия;

поисковая стратегия;

поиск информации в РИС ГВС;

послепоисковая стратегия (обработка релевантной информации).

Такая последовательность обусловлена системностью подхода к поиску информации в РИС ГВС, позволяющего выявить в ходе информационной работы наиболее эффективные лексические единицы, и на их основе реализовать информационную потребность. На практике возможно выполнение всех приведенных этапов одномоментно, но за счет хаотичности (многоитерационности) будет достаточно сложно контролировать процесс поиска и обработки информации. Именно последовательное прохождение всего жизненного цикла поиска и обработки информации в РИС ГВС с возвращением на предыдущие этапы в случае обнаружения новых категорий (информационных параметров), влияющих на информационную работу, позволяет сократить совокупное время, затраченное на поисковую и аналитическую работу, а также приводит к удовлетворению информационной потребности пользователя.

В рамках предпоисковой стратегии формируются лексические единицы, составляющие пользовательский тезаурус и лежащие в основе создания поисковых предписаний. Лингвистической основой тезауруса [2] являются три базовые группы лексических единиц: ключевые слова и дескрипторы предметной области, а также существующие системы классификаций, а также специфические лексические единицы, такие как авторы, держатели информации, патентообладатели и другие. На этой основе формируется тезаурус всех групп лексических единиц, включающий:

- эталонные дескрипторы;
- дескрипторы и ключевые слова, расширяющие предметную область (на основе существующих и специализированных систем классификации);
- дескрипторы и ключевые слова, суживающие предметную область (на основе существующих и специализированных систем классификации);
- ключевые слова и словоформы;
- ключевые слова и синонимы;
- ключевые слова и антонимы;
- дескрипторы синонимии;
- дескрипторы антиномии;
- дескрипторы антонимии;
- дескрипторы с учетом расширенной предметной области;
- дескрипторы с учетом сужения предметной области;
- другие.

Аналитической основой формирование тезауруса являются имитационные модели синонимии, антиномии и антонимии, а также индуктивные и дедуктивные модели. Синонимия [3] применяется для расширения границ информационной потребности и позволяет выполнять поиск и обработку информации в смежных предметных областях. Антонимия, наряду с синонимий, также применяется для расширения границ информационной потребности, но имеет другой очень важный аналитический аспект, позволяющий осуществлять информационную работу в противоположных предметных областях с сохранением её лингвистического смысла. Антиномия [4, 5] необходима для сужения границ информационной потребности за счет установления четких разграничений между лингвистическими и семантическими аспектами организации лексических единиц дескрипторного типа без потери связности искомой информации и в итоге ее смысла.

Формирование предметной области, как одной из групп лексических единиц, основывается на общеизвестных (ГРНТИ, УДК, МПК, ББК, ОКПД2 и пр.) и специализированных системах классификации и рубрицирования. Применение таких инструментов при поиске и обработке информации в РИС ГВС обусловлено их информационной структурой, которая представляет собой набор жестко связанных по смыслу между собой дескрипторов и ключевых слов. Для расширения предметной области используется дедуктивный метод [6] – от общего к частному, для сужения предметной области индуктивный метод [7] – от частного к общему.

Поисковая стратегия нацелена на подготовку поисковых предписаний из массива лексических единиц пользовательского тезауруса. Для формирования поисковых предписаний используется весовая мате-

матическая модель с назначением веса от 0,1 до 1,0 каждой группе лексических единиц, в том числе для повышения точности и значимости отдельных лексических единиц возможно назначение весов отдельным лексическим единицам внутри каждой группы. Весовое распределение осуществляется пользователем (оператором) самостоятельно, исходя из значимости (смысла) каждой группы применяемых лексических единиц. Как правило, это распределяется следующим образом[1]: ключевые слова (КС) – 1; эталонные дескрипторы ($Д_Э$) – 1; предметная область (ПО) – 0,9; дескрипторы синонимы ($Д_С$) – 0,8; дескрипторы с учетом расширения ($Д_{ПОр}$) и с учетом сужения предметной области ($Д_{ПОс}$) – 0,6; расширение предметной области ($ПО_Р$) – 0,4; дескрипторы антиномии ($Д_А$) – 0,3; дескрипторы антонимии ($Д_{АТ}$) – 0,2; словоформы ключевых слов ($КС_{СЛ}$) – 0,1 и так далее. Такая конструкция необходима не только для формирования набора лексических единиц в поисковых предписаниях, но и для расстановки лексических единиц в поисковом предписании относительно друг друга. Цель поиска информации в любом текстовом документе – это найти не только просто встречаемость лексических единиц в тексте (релевантность), но и последовательность их встречаемости (пертинентность), что даст возможность осуществлять смысловой поиск информации, так как смысл определяется не конкретными словами, а их последовательностью.

Конструкции поисковых предписаний (ПП) могут выглядеть следующим образом:

$ПП_1$ – Дэ1 (1,0) + КС1 (1) + ПО1 (0,9);
$ПП_2$ – КС2 (1) + КС3 (1) + $ПО_Р$ (0.4);
$ПП_3$ – Дэ2 (1) + $Д_С$ (0.8) + $ПО_Р$ (0,4) + КС4 (0,3);
$ПП_N$ – $N_1 + N_2 + N_3 + N_N$.

Ограничения на формирование поисковых предписаний оказывают применяемые механизмы поиска информации в различных структурированных и неструктурированных источниках информации РИС ГВС. К таким ограничениям следует отнести сложность применяемых методов поиска информации в РИС ГВС, количество символов поисковой строки и др. Структурированные источники информации позволяют обрабатывать, как правило, простые поисковые предписания, состоящие из одиночных лексических единиц, или ряда поисковых переменных, к ним относятся ключевые слова и одиночные несложные дескрипторы. Неструктурированные источники информации РИС ГВС имеют более продвинутые алгоритмы поисковых механизмов (например, *Google Hummingbird*), для них характерно использование всех возможных групп лексических единиц, в большей степени дескрипторного типа. Единственная проблема, которая стоит перед любым поисковым предписанием, как бы оно изначально ни казалось отвечающим всем аспектам информационной потребности, – это его заранее не известная эффективность, которую возможно проверить только в процессе поиска и обработки информации, и только уже потом делать выводы о его состоятельности, именно для этого и необходим многоитерационный процесс поиска, чтобы понять (выявить) эффективные конструкции поисковых предписаний.

Результатом разработки поисковой стратегии является подготовленный массив поисковых предписаний, отвечающий и определяющий информационную потребность пользователя (оператора).

Поиск информации в РИС ГВС проводится многоитерационно с применением существующих поисковых машин РИС ГВС. Чем больше проведенных итераций, тем лучше будут сужены границы информационной потребности. Для установления в процессе поиска информации семантических связей между информационной потребностью пользователя (оператора) и найденными релевантными документами необходимо применять интеграционную модель смыслового поиска информации в РИС ГВС [8], состоящую из следующих этапов формирования:

эталонного поискового образа документа ($ПОД_Э$);
поискового образа запроса (ПОЗ) [9];
расширенного поискового образа запроса (РПОЗ);
поискового образа документа найденной релевантной информации ($ПОД_Р$).

В процессе поиска информации происходит оценка применения сформированных поисковых предписаний, на основе которых выявляются наиболее значимые комбинации лексических единиц, описывающих наиболее точно информационную потребность. В случае неполучения результатов от конкретного поискового предписания, рекомендуется возвратиться на предыдущий этап «поисковая стратегия» и переформулировать и/или перегруппировать поисковое предписание, нерезультативные поисковые предписания исключить из массива и далее не применять их для поиска информации в РИС ГВС. При обнаружении новых, ранее не использовавшихся информационных параметров, таких как новое направление предметной области, новые ключевые слова, дескрипторы, авторы, держатели и патентообладатели, следует вернуться на ступень «анализ информационного пространства», а в случае других параметров – на этап «предпоисковая стратегия» и продолжить информационную работу с этого этапа, т.е. с самого начала.

По итогам поиска информации в РИС ГВС получаем релевантный массив информации, соответствующей информационному (поисковому) запросу.

Обработка релевантной информации (послепоисковая стратегия) проводится исключительно аналитическим путем непосредственно пользователем, поставившем задачу реализации информационной потребности. Такое положение обусловлено тем, что только специалист способен придать смысл данным [10] и получить на выходе пертинентную информацию, соответствующую информационной потребности.

Для поддержки аналитической работы применяется интеграционная модель ранжирования и реферирования найденных (релевантных) документов, которая используется для правильной классификации и определения точности вхождения (совпадения) $ПОД_Р$ в $ПОД_Э$ с учетом вероятностной математической модели (основанной на частоте повторения слов и подсчете средних частот употребления слов в документе), являющейся основой для формирования частотного

[1] Распределение весовых коэффициентов по группам лексических единиц определено нами исходя из своего личного практического опыта работы с информацией в РИС ГВС.

словаря документа. В математическую модель заложен единственный семантический параметр (критерий) определяющий точность вхождения $ПОД_Р$ в $ПОД_Э$ – доля (процент) вхождения (совпадения) искомой и найденной информации. В зависимости от информационной потребности пользователь (оператор) самостоятельно настраивает необходимые долевые границы (пороги) вхождения $ПОД_Р$ в $ПОД_Э$. Реферирование найденных документов необходимо для преобразования документальной информации в сокращенный машиночитаемый вид, обеспечивающий семантически адекватное изложение содержания первичного документа, на основании которого пользователь (оператор) принимает решения о пертинентности найденного документа.

В результате аналитической обработки первичных документов, отобранных из РИС ГВС, формируется массив пертинентных документов, соответствующих информационной потребности пользователя (оператора).

ПРАКТИЧЕСКАЯ ОСНОВА РЕАЛИЗАЦИИ КОНЦЕПТУАЛЬНЫХ ПРЕДЛОЖЕНИЙ

Практическая основа реализации концептуальных предложений основана на разработке системы (сервиса) поиска и семантической (смысловой) обработки информации в РИС ГВС (далее – Система). По своей сути Система относится к классу экспертно-аналитических систем с возможностью формирования базы знаний, исходя из предметной области и поведения пользователя.

Система нацелена на поддержку жизненного цикла информационной работы пользователя (оператора) с общедоступной (открытой) структурированной и неструктурированной информацией и удовлетворение информационной потребности пользователя за счет создания инструментов (в том числе интерактивных) поддержки на каждом этапе поиска и обработки информации с формированием самоорганизующейся базы знаний в виде настраиваемой онтологии предметной области с учетом предпочтений пользователя.

В Систему заложен принцип многоитерационного взаимодействия пользователя (оператора) с Системой в процессе автоматизированного, а не автоматического, как в существующих поисковых системах, поиска и обработки информации на каждом этапе, начиная с анализа информационной структуры объекта исследований, определения терминологической структуры, формирования поисковых предписаний, поиска информации, обработки результатов поиска и сопоставления их с информационными задачами, заканчивая формированием отчетных материалов.

Система может быть использована для работы:

- со всеми видами информации при решении задач профессиональной деятельности в рамках информационного обслуживания и информационно-аналитической поддержки производственных процессов;
- с научно-технической, патентной и другой информацией для принятия управленческих решений по развитию объектов исследований в рамках жизненного цикла продукции, особенно на этапах НИР, ОКР;
- с маркетинговой, деловой и другой информацией в рамках принятия решений по продвижению продукции на рынки сбыта;
- с оперативной новостной информацией (социальной, экономической, политической) для понимания действительной картины и большего охвата информационного пространства мира и отдельных территорий;
- со всеми видами информации для решения задач по отдельным аспектам государственного развития в рамках деятельности органов государственной власти;
- других видов деятельности.

Система обеспечивает:

- формирование интерактивных сервисов (инструментов) для поддержки поиска и обработки информации в виде графически-ориентированного конструктора, позволяющего пользователю настраивать Систему под свои информационные потребности;
- формирование и планирование (план-график) заданий (от 1 до *N*) на поиск и обработку информации в РИС ГВС, позволяющих оперативно в режиме реального времени осуществлять поиск и обработку информации в соответствии с поставленными задачами;
- создание единой платформы поиска информации в РИС ГВС, позволяющей осуществлять поиск по выбранным доступным механизмам, таким как современные поисковые машины, различные поисковые сервисы электронных ресурсов и баз данных консолидированных (структурированных) информационных ресурсов научно-технической, патентной и другой информации;
- формирование иерархической терминологической структуры (дерева/онтологии) предметной области с использованием существующих общеизвестных (УДК, МПК, СПК, ГРНТИ, ББК и др.) и специализированных систем классификации по отдельным предметным областям и даже технологическим направлениям;
- возможности создания иерархической структуры онтологии предметной области непосредственно пользователем под свои информационные потребности и использование её для поиска и обработки информации;
- управление разработкой тезауруса предметной области (в виде интерактивного дерева), как в ручном, так и в автоматизированном режиме, с применением методов индукции и дедукции, а также синтаксического анализа;
- изучение внешнего информационного пространства предметной области с применением методов синтаксического и лексического анализа для обработки информации и извлечения из неё лексических единиц, подпадающих под информационную потребность пользователя;
- применение методов синонимии, антономии и антиномии для расширения, сужения и определения границ предметной области на основе общеизвестных и специализированных систем классификации;
- формирование системы поисковых предписаний на основе адекватной математической модели с применением методов весовых коэффициентов;

- создание единой мультиязычной поисковой платформы поиска информации на различных языках для более полного охвата мирового информационного пространства с применением настраиваемых и встраиваемых средств автоматического машинного перевода;
- возможности формирования пользовательской базы данных (перечня) структурированных и неструктурированных источников информации;
- формирование единой настраиваемой системы обработки информации на основе реферирования найденных документов (материалов) с применение рейтинговой математической модели в соответствии с информационной потребностью пользователя;
- управление и настройку процесса обработки и выдачи информации на основе концепции математической модели с внешним управляющим возмущением, где результат обрабатывается и выдается в соответствии с заданными параметрами (настраивается пользователем);
- формирование отчетных документов в автоматизированном режиме для определенных видов информационно-аналитических исследований (патентных, маркетинговых и др.).

Работы по поиску и семантической обработке информации в РИС ГВС Система реализует в следующих режимах:

- автоматическом – пользователь вводит базовые (основные) параметры, необходимые для идентификации информационной потребности, далее Система выполняет всю процедуру в автоматическом режиме, с настроенными параметрами (низкая / непредсказуемая точность поиска);
- автоматизированном – пользователь вводит базовые (основные) параметры, необходимые для идентификации информационной потребности, далее Система поэтапно начинает выполнять заложенную процедуру, но при этом на каждом этапе просит пользователя верифицировать результаты выполнения этапа с возможностью внести определенные пользовательские корректировки (наиболее предпочтительный режим работы);
- ручном – пользователь проводит всю процедуру самостоятельно, Система только дает последовательность этапов, которые нужно выполнить; пользователь может самостоятельно выбирать этапы поиска и обработки информации (необходимо владеть технологиями поиска информации).

Система специализируется на поиске и семантической обработке: научно-технической; патентной; деловой и маркетинговой; новостной и любой другой информации.

В принципе Система может работать с любым видом информации, но для более точного поиска и обработки необходимо понимать информационную (лингвистическую) природу различных видов информации. Система не направлена на решение поисковых задач, возникающих в повседневных жизненных ситуациях пользователей, такие задачи решаются другими средствами поиска (например, поисковыми машинами РИС ГВС и др.), она ориентирована на потребителей (пользователей), профессионально занимающихся информационно-аналитической работой: сотрудников специальных служб и ведомств, аналитиков и специалистов государственных органов власти, аналитиков информационно-консалтинговых агентств и бюро, журналистов и аналитиков СМИ, специалистов подразделений маркетинга и продаж, инженеров, конструкторов и технологов научно-технологических и конструкторских подразделений промышленных предприятий, сотрудников научно-исследовательских институтов, ученых, исследователей и изобретателей и др.

В Системе осуществляется поддержка функциональных возможностей на предпоисковой, поисковой и послепоисковой стадиях поиска и семантической обработки информации в РИС ГВС.

Предпоисковая стадия – в общем виде проводятся работы по изучению (анализу) информационной структуры объекта исследований, создается информационное пространство информационной потребности, строится терминологическая структура, формируются поисковые предписания. На данной стадии реализуется следующие основные функциональные возможности, разделенные на два этапа (предпоисковая стратегия и поисковая стратегия):

- формирование конструктора заданий, позволяющего в интерактивном виде управлять (создавать, редактировать, удалять) заданием на поиск и обработку информации в РИС ГВС, с указанием основных параметров выполнения задания (время, график, объемы информации и т.п.);
- построение в интерактивном виде с помощью блочно-графического конструктора информационного пространства объекта и предмета информационной потребности;
- применение интерактивного инструмента, позволяющего с использованием графической структуры информационного пространства информационной потребности создавать лексические единицы (дескрипторы, ключевые слова, предметная область, и др.);
- автоматическое построение дескрипторов из произвольного ряда ключевых слов;
- формирование терминологической структуры групп лексических единиц в соответствии с информационными потребностями;
- расширение и сужение предметной области объекта исследований с учетом существующих общеизвестных и специализированных систем классификации (установка градаций полноты);
- разработка словоформ, синонимов, антонимов и т.п. ключевых слов с применением промежуточного сервиса, подключаемого к существующим информационным ресурсам, позволяющим обеспечивать выполнение данного функционала (установка градаций полноты);
- построение в интерактивном виде с помощью блочно-графического конструктора дерева иерархической структуры синонимии, антиномии, антонимии и т.п.;
- создание в блочно-графическом виде информационной иерархической структуры дерева онтологии информационного пространства информационной потребности, с возможностью изменения ветвей и листьев дерева онтологии;
- выбор других языков для поиска информации и автоматический (машинный) или ручной (пользо-

вательский) перевод терминологической структуры информационной потребности (для другого языка также доступен весь функционал Системы);

- создание в виде конструктора сервиса мультиязычных специализированных систем классификации, позволяющих устанавливать прямые связи используемой терминологии на разных языках;
- формирование: $ПОД_Э$ с возможностью внесения изменений в структуру $ПОД_Э$ (доступны автоматический и полуавтоматический режимы создания), ПОЗ с возможностью внесения изменений в структуру ПОЗ (доступны автоматический и полуавтоматический режимы создания), РПОЗ с возможностью внесения изменений в структуру РПОЗ (доступны автоматический и полуавтоматический режимы создания), поисковых предписаний, с учетом специфики поисковых механизмов информационных ресурсов, с применением математической модели с весовыми коэффициентами, отображающими степень важности лексических единиц информационной потребности;
- построение дерева (онтологии) сформированных поисковых предписаний в виде блочно-графического конструктора с возможностью внесения изменений;
- установка степени (доли) вхождения найденного материала в предметную область информационной потребности объекта исследований;
- запись статистических данных о результативности поисковых предписаний, групп лексических единиц и отдельных лексических единиц (выделение цветом, указание доли результативности и т.п.).

Поисковая стадия – реализуется через промежуточные сервисы (синтаксические анализаторы) подключения к существующим поисковым системами РИС ГВС (поисковые машины: *Google*, *Yandex*, *Yahoo*, *Bing* и т.п.), структурированным источникам (журналы, базы данных, библиотеки и т.п.) научно-технической и патентной информации (электронные ресурсы) открытого доступа (*open access/source*). На данной стадии реализуются следующие основные функциональные возможности Системы (поиск структурированной и неструктурированной информации):

- формирование блочно-графического конструктора, позволяющего выбирать необходимые источники структурированной и неструктурированной информации;
- поиск информации в структурированных источниках, с применением промежуточного сервиса (синтаксического анализатора) обработке поисковых инструментов ресурсов, учитывающего особенности механизмов поиска;
- поиск информации в неструктурированных источниках с применением промежуточного сервиса (синтаксического анализатора) обработки поисковых инструментов ресурсов, учитывающего особенности механизмов поиска;
- синтаксический анализ информации в соответствии с терминологической структурой информационной потребности объекта исследований;
- скачивание информации, попадающей под установленную степень информационной потребности в базу данных для дальнейшей обработки;
- запись статистических данных о поиске для определения результативности того или иного поискового механизма Системы.

Послепоисковая стадия – первичная обработка и анализ информации, в том числе:

- формирование для каждого отобранного информационного материала поискового образа документа ($ПОД_Р$);
- ранжирование отобранных $ПОД_Р$ в соответствии с установленной степенью (долей) вхождения $ПОД_Р$ в информационную потребность объекта исследований, по убыванию, начиная самого высокого;
- построение дайджеста отобранной информации, позволяющего в кратком виде отобразить смысл информационного материала с возможностью перейти на просмотр полной версии;
- определение вида документов (научные статьи, описания изобретений, новости и т.п.), синтаксический анализ структуры информации (заголовок, реферат/аннотация, библиографические данные, основной текст, список литературы, и т.п.), расстановка тегов по тексту в зависимости от целей обработки;
- лексический анализ информационной структуры отобранных документов на предмет выявления групп лексических единиц (ключевые слова, дескрипторы, авторы, разработчики, держатели и т.п.), отображающих информационную потребность пользователя;
- создание отчетных документов по информационно-аналитическим исследованиям (патентные, маркетинговые) с применением синтаксического анализа информационной структуры отобранных документов.

ЗАКЛЮЧЕНИЕ

Предложенный в настоящей статье концептуальный подход к поиску и семантической обработке научно-технической информации в распределенных системах Интернета позволит:

- создать методологические основы для понимания лингвистической и семантической природы информации, обеспечить поддержку всех этапов жизненного цикла поиска и семантической обработки общедоступной (открытой) структурированной и неструктурированной информации в РИС ГВС;
- снизить общее совокупное время, затрачиваемое на поиск и семантическую обработку информации в РИС ГВС;
- повысит контроль и границы полноты охвата информационных ресурсов РИС ГВС;
- обеспечить контроль достоверности, получаемой из РИС ГВС информации.

С практической точки зрения концептуальный подход создаст фундаментальные основы для развития решений, направленных на поддержку процессов поиска и семантической обработки информации в РИС ГВС, а также практические условия по решению проблем управления внешней (дополнительной) информацией, в частности:

- выработает предпосылки для развития единой платформы смысловой (семантической) обработки открытой (общедоступной) структурированной и неструктурированной информации;

- сформирует единую систему организационно-информационной поддержки технологии (постановки задач) поиска и семантической обработки информации для специалистов, профессионально занимающихся вопросами проведения информационных и аналитических работ;
- построит структуру единой самоорганизующейся базы знаний (экспертных знаний), ориентированную на информационную потребность пользователя и позволяющую в режиме реального времени осуществлять информационную (аналитическую) работу на постоянной основе;
- сформирует функциональные основы для разработки семантического поискового робота (*semantic search engine*), позволяющего в режиме реального времени обрабатывать большие массивы неформальной внешней (дополнительной) структурированной и неструктурированной информации, размещенной в РИС ГВС.

СПИСОК ЛИТЕРАТУРЫ

1. Трусов В.А., Трусов А.В. Модель поиска информации в распределенных информационных системах сети Интернет // Научно–техническая информация. Сер. 2. – 2011. –№ 8. – С. 29–31.
2. Трусов В.А. Построение тезаурусов, тематических классификаций и рубрикаторов для поиска информации в распределенных информационных системах // Информационные ресурсы России. – 2011. – № 3(121). – С. 9-13.
3. Чешко Л.А. О синонимах и словаре синонимов русского языка. – Москва, 1967. – 96 с.
4. Гийом Г. Принципы теоретической лингвистики. – Москва, 1992. – 224 с.
5. Радзиевская Т.В. Прагматические противоречия при текстообразовании // Логический анализ языка. – Москва, 1990. – С.148-162.
6. Пятницын Б.Н. Индуктивная логика и формирование научного знания. – Москва: Наука, 1987. – 175 с.
7. Финн В.К. Синтез познавательных процедур и проблема индукции // Научно-техническая информация. – Сер. 2. – 1999. – № 1/2. – С. 8–45.
8. Трусов В.А., Трусов А.В. Подходы к формированию смыслового поиска информации в распределенных информационных системах сети интернет // Информационные ресурсы России. – 2011. – № 2(120). – С. 20–24.
9. Трусов В.А. Модель построения поискового образа запроса в распределенных информационных системах сети интернет / В.А. Трусов // Научно–техническая информация. Сер. 2. – 2011. – № 5. – С. 18–22.
10. Гиляревский Р.С. Основы информатики: курс лекций. – Москва: Экзамен, 2004. – 320 с.

Материал поступил в редакцию 07.03.21.

Сведения об авторе

ТРУСОВ Владимир Александрович – кандидат технических наук, доцент, начальник отдела Пермского ЦНТИ – филиала ФГБУ «РЭА» Минэнерго России; доцент Пермского национального исследовательского политехнического университета
e-mail: tva@permcnti.ru

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

УДК 004.89:616.89

Е.Ф. Фабрикантова, Д.В. Романов

Применение ДСМ-метода автоматизированной поддержки исследований в области психиатрии*

Приводятся описание ДСМ-метода и определение интеллектуальной системы типа ДСМ (ИС-ДСМ), реализующей этот метод. Обсуждается вопрос, почему ДСМ-метод является методом интеллектуального анализа данных, а ИС-ДСМ – интеллектуальной системой. Рассматриваются требования ДСМ-метода к представлению данных для анализа с его помощью. Приводятся примеры исследований из психиатрии и междисциплинарных исследований. На их основе формулируются некоторые принципы подготовки данных для анализа с помощью ИС-ДСМ.

Ключевые слова: *психиатрия, психометрическая (шкальная) оценка, ДСМ-метод автоматизированной поддержки исследований, интеллектуальный анализ данных с помощью ДСМ-метода, интеллектуальная система типа ДСМ (ИС-ДСМ), представление данных для анализа с помощью ИС-ДСМ*

DOI: 10.36535/0548-0027-2021-04-2

ВВЕДЕНИЕ

Накопленные в психиатрии экспертные знания (обобщения клинических данных) для диагностики психических расстройств содержатся в виде рубрик (синдромов и симптомов расстройств) в классификациях МКБ-10 и DSM-5. Совершенствование диагностики и лечения психических расстройств требует как проведения исследований в самой психиатрии, так и привлечения данных из других областей, например, психологии, генетики, нейронаук, т. е. междисциплинарных исследований. Цель таких исследований – извлечение новых знаний из данных, создание эмпирической теории изучаемого явления.

В Диагностическом и статистическом руководстве по психическим расстройствам, 5-е издание (DSM-5 – *Diagnostic and Statistical Manual of mental disorders*, *fifth edition*) [1] введено определение психического расстройства, которое обозначено как «синдром, характеризующийся клинически значимым нарушением регуляции когнитивного функционирования и эмоций человека или поведения, который отражает дисфункцию психологических, биологических и онтогенетических процессов, лежащих в основе психического функционирования. Психические расстройства обычно связаны со значительным дистрессом или ограничением в социальной, профессиональной или других важных сферах деятельности. Ожидаемые или культурально приемлемые реакции на обычные стрессовые события или утраты, такие как смерть любимого человека, не является психическим расстройством» [2].

При проведении исследований в психиатрии это определение лежит в русле стремления к отражению целостности организма, его интегральности, например, добавляя биологические факторы: генетические, нейрофизиологические, биохимические, иммунные и др. Актуальны также междисциплинарные исследования, например, в области психосоматических и психосоциальных расстройств, учитывающие взаимодействия организма со средой и социумом. Добавим еще, что развитие информационных технологий и появление новых технических средств, например, магнитно-резонансной томографии (МРТ), обогащает эти исследования нейровизуализационными данными[1]. Такие работы не могут быть проведены без применения компьютера: необходимо учесть слишком много факторов, которые эксперт не может охватить для самостоятельного формирования ассоциаций, правил, закономерностей.

* Работа выполнена при частичной финансовой поддержке РФФИ (проект № 18-29-03063, проект № 19-07-01119).

[1] Заметим, что, несмотря на тенденцию к увеличению объема биологических данных, диагноз психического расстройства в настоящее время не может быть установлен только на их основе, а до сих пор верифицируется клинически.

В психиатрии нет собственных формальных средств для анализа эмпирических данных, поэтому в ней используются статистические методы анализа. Если исходить из того, что целью всякого исследования в эмпирической предметной области (а психиатрия именно такая предметная область) является создание эмпирической теории изучаемого явления, то ответом на научные работы в области психиатрии следует считать современное направление в искусственном интеллекте – обнаружение новых знаний в эмпирических данных, интеллектуальный анализ данных. Такие возможности реализованы в ДСМ-методе автоматизированной поддержки исследований, принципы которого разработал профессор В.К. Финн [3], и в интеллектуальной системе типа ДСМ (ИС-ДСМ), реализующей этот метод. Формальным средством анализа эмпирических данных, рассматриваемым в предлагаемой работе, является ДСМ-рассуждение.

ДСМ-метод и ИС-ДСМ неоднократно использовались для решения задач в различных областях медицины [см., например, 4, 5]. Настоящая работа посвящена использованию ДСМ-метода и ИС-ДСМ для исследований психиатрических, психосоматических и психосоциальных расстройств, т.е. в новой для них области – психиатрии. Как видно из приведенного определения психического расстройства, в поле изучения психиатрии сегодня находятся исследования, рассматривающие взаимодействие патологического поведения пациента с соматическими заболеваниями, генетикой, физиологией и т.д., с одной стороны, а с другой, – с социальными и психологическими аспектами поведения человека в норме и патологии. Эти исследования являются междисциплинарными – современная психиатрия расширила свои границы, вышла за пределы «классического» психопатологического подхода. Для проведения таких работ объединяются субъективные (психиатрические, психологические, социологические) и объективные (соматические) данные. ДСМ-метод и ИС-ДСМ не различают природы данных, однако предъявляют некоторые требования к их представлению для анализа.

Приведем несколько примеров представления данных, связанных с требованиями ДСМ-метода и ИС-ДСМ для их анализа. Это исследование психосоматического (функционального) зуда и аффективной патологии, установление связи некоторых генетических характеристик и эмоционального и/или волевого дефицита у пациентов с негативным синдромом при расстройствах шизофренического спектра, а также депрессии, тревоги, стигматизации и расстройства образа тела (дисморфического расстройства) у пациентов с кожными заболеваниями. При этом покажем «идеальные» постановки исследований (как они сформулированы экспертами) и постановки, обеспечивающие выполнение экспериментов на ИС-ДСМ за приемлемое время. Дело в том, что ДСМ-метод является логико-комбинаторным, и алгоритмы, реализующие его, переборные. Шкалы же, используемые для оценки психического состояния пациентов, содержат признаки, имеющие большое количество значений, что приводит к значительным временным затратам при экспериментах с применением ЭВМ.

I. ЗНАНИЯ В ПСИХИАТРИИ

1. Диагностика: классификационные системы

Для диагностики психических расстройств на основе клинических данных и назначения лечения врачи-психиатры используют знания, клиническое мышление, опыт и интуицию. Кроме того, накоплено много обобщений клинических данных – экспертных знаний о психических расстройствах – это синдромы и симптомы заболеваний, некоторые совокупности «ядерных» клинических данных, характерные для проявления болезни пациентов при конкретном психическом расстройстве. Это отражено, например, в МКБ-10 (Международная классификация болезней 10-го пересмотра) [6] и номенклатуре психических расстройств DSM-5 (*Diagnostic and Statistical Manual of mental disorders*, *fifth edition* – Диагностическое и статистическое руководство по психическим расстройствам, 5-е издание) [1].

Классификационные системы, представленные в МКБ-10 и DSM-5, созданы экспертами Всемирной организации здравоохранения (ВОЗ), национальными ассоциациями психиатров и т.д. (ср. «приобретение знаний» для экспертных систем в искусственном интеллекте) и отражают результаты анализа данных с помощью статистических методов.

2. Современные клинические (субъективные) эмпирические данные в психиатрии – формализованные психометрические методики

Для объективизации симптомов психических расстройств и помощи в их диагностике используются методы дополнительного обследования – психометрические методики (шкалы) [7]. Приведем в качестве примера шкалу Гамильтона для оценки тяжести депрессии (*Hamilton Rating Scale for Depression* – *HDRS*) – клиническое пособие, разработанное в 1960 году Максом Гамильтоном (университет Лидса, Великобритания), для количественной оценки состояния пациентов с депрессивными расстройствами.

Шкала заполняется врачом и содержит субъективные оценки психического состояния пациента, затем баллы суммируются, что позволяет ранжировать депрессию по степени тяжести: 0-6 – норма, 7-17 – легкое депрессивное расстройство; 18-24 – депрессивное расстройство средней степени тяжести; 25 и выше – депрессивное расстройство тяжелой степени.

Оценка по шкале Гамильтона позволяет не только верифицировать клинический диагноз, установленный с помощью диагностических критериев МКБ-10 или DSM-5, но и стандартизировать и объективизировать диагноз, а также «отсечь» случаи, в которых депрессивная симптоматика не достигает диагностического порога депрессии либо отсутствует. Кроме того, результатом такой оценки является «профиль» депрессии, отражающий наличие/выраженность тех или иных симптомов депрессии из представленного перечня.

Шкалы создавались для статистического анализа данных и имеют большое значение для стандартизации и объективизации знаний в психиатрии, а также для доказательства эффективности лечения.

Таблица 1

Шкала Гамильтона для оценки тяжести депрессии

№	Симптом	Оценка в баллах
1	Пониженное настроение (переживания печали, безнадежности, собственной беспомощности и малоценности)	От 0 до 4
2	Чувство вины	От 0 до 4
3	Суицидальные тенденции	От 0 до 4
4	Ранняя бессонница (трудности при засыпании)	От 0 до 2
5	Средняя бессонница	От 0 до 2
6	Поздняя бессонница	От до 2
7	Работоспособность и активность (работа и деятельность)	От 0 до 4
8	Заторможенность	От 0 до 4
9	Ажитация (возбуждение)	От 0 до 4
10	Тревога психическая	От 0 до 4
11	Тревога соматическая	От 0 до 4
12	Желудочно-кишечные соматические нарушения (симптомы)	От 0 до 2
13	Общесоматические симптомы	От 0 до 2
14	Расстройства сексуальной сферы (генитальные симптомы)	От 0 до 2
15	Ипохондрические расстройства (ипохондрия)	От 0 до 4
16	Потеря веса	От 0 до 4
17	Отношение к своему заболеванию (критичность отношения к болезни)	От 0 до2

II. ДСМ-МЕТОД И ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА ТИПА ДСМ (ИС-ДСМ)

Первоначально ДСМ-метод имел название «ДСМ-метод порождения гипотез». Современное его название – «ДСМ-метод автоматизированной поддержки исследований». ДСМ-метод и интеллектуальная система (ИС-ДСМ), реализующая этот метод, имеют большой опыт использования для интеллектуального анализа данных в фармакологии, социологии, медицине, криминалистике [8].

Основным понятием в ДСМ-методе является ДСМ-рассуждение, в котором реализуется синтез (взаимодействие) трех познавательных процедур: индукция + аналогия + абдукция.

ДСМ-рассуждение применяется к данным, организованным в два массива объектов, – базу положительных и базу отрицательных примеров (базы фактов) исследуемого эффекта (эффектов). В них содержатся описания исследуемых объектов. В рассматриваемых в настоящей работе исследованиях объектами являются пациенты и здоровые (с точки зрения проводимого исследования) люди. В базах фактов объекты представляются на языке описания объектов, в ДСМ-парадигме – это совокупность дифференциальных признаков с их значениями. Признаками являются симптомы психических расстройств, описания генетики, психологии, соматических заболеваний и др.

Основным результатом применения ДСМ-рассуждения являются гипотезы о причинах, связях в данных между эффектом (и его отсутствием) и комплексами значений признаков, описывающих объекты.

Процедура индукции (анализ данных), обобщения данных. На этом этапе порождаются все возможные сходства объектов (операция сходства выполняется между всеми возможными комбинациями объектов – по два, по три, по четыре и т. д.) отдельно для положительных и отрицательных примеров эффекта. Сходство определяется через общность значений одних и тех же признаков в двух или более примерах. Найденные по всем объектам (с учетом некоторых условий, фильтров) сходства в терминологии ДСМ-метода интерпретируются как гипотезы о причинах изучаемого эффекта (эффектов). Критерии (фильтры) для отбора гипотез: количество примеров, которые участвуют в порождении сходства, – так называемых «родителей», – должно превышать некоторый выбранный порог, т. е. отбираемое сходство должно встречаться у достаточно большого числа объектов исходного массива (эти значения чаще всего подбираются экспериментально); сходство как набор признаков не должно одновременно встречаться в положительных и отрицательных примерах (запрет на контрпример).

Процедура аналогии (предсказание, доопределение). На этом этапе происходит предсказание эффекта для примеров из специальной базы объектов, наличие или отсутствие эффекта у которых неизвестно (τ-примеры). Предсказание выполняется с помощью причин, полученных на этапе индукции: если в описание τ-объекта входят положительные гипотезы о причинах, то он доопределяется как положительный пример эффекта, если входят отрицательные гипотезы, то объект доопределяется как отрицательный. Если в описание τ-объекта входят и положительные, и отрицательные гипотезы, то он объявляется как объект с «противоречивым предсказанием», если в описание не вошли ни положительные, ни отрицательные гипотезы, то объект сохраняет свою неопределенность и остается τ-примером.

Процедура абдукции (принятие гипотез посредством объяснения исходных фактов). На этом этапе проверяется объяснимость исходного массива полученными гипотезами. Этот показатель вычисляется для каждой гипотезы как отношение числа фактов (примеров), из которых она была получена (число «родителей» гипотезы), к общему числу примеров. Обычно он выражается в процентах.

В первой версии ДСМ-метода использовалось однократное применение ДСМ-рассуждения. В новой версии метода, в которой ДСМ-рассуждение применяется к последовательности расширяющихся баз положительных и отрицательных примеров исследуемого эффекта, образуя ДСМ-исследования, результатом которых должно быть обнаружение эмпирических закономерностей (регулярностей) в данных [9, 10].

Таким образом, сейчас

ДСМ-анализ = ДСМ-рассуждение +
+ДСМ-исследование.

1. Знания в интеллектуальной системе типа ДСМ [11]

(1) знания нулевого уровня (знания$_0$): элементы базы фактов (БФ), где факт есть элементарное высказывание с типами истинностных значений «1» (фактически истинно), «-1» (фактически ложно), «τ» (неопределенно);

(2) знания первого уровня (знания$_1$): логические комбинации знаний нулевого уровня;

(3) знания второго уровня (знания$_2$): представление процедур (процедурное знание) и гипотезы, полученные применением процедур;

(4) знания третьего уровня (знания$_3$): аксиомы квазиаксиоматических теорий (КАТ) – дескриптивные аксиомы и аксиомы структуры данных;

(5) знания четвертого уровня (знания$_4$): обнаруженные эмпирические закономерности (ЭЗК);

(6) база знаний (БЗ) есть множество знаний$_i$, где $i = 0, 1, 2, 3, 4$.

База знаний образует систему знаний в интеллектуальной системе (ИС), представимую посредством КАТ, где КАТ $\mathbf{T} = \langle \Sigma, \Sigma', \mathbf{R} \rangle$, Σ – аксиомы, Σ' – множество фактов и гипотез, R – множество правил правдоподобного и достоверного вывода.

Интеллектуальная система типа ДСМ (ИС-ДСМ) реализует ДСМ-метод автоматизированной поддержки исследований.

Как программный комплекс ИС-ДСМ состоит из модулей Решателя, реализующих методы (прямой, обратный, обобщенный, ситуационный и др.); стратегии (например, метод для положительных примеров + метод для отрицательных примеров + возможные условия (например, условие «запрета на контрпример») и др.); БФ – баз положительных и отрицательных фактов эффекта; БЗ – базы знаний (порождаемые гипотезы) и интерфейса:

ИС-ДСМ = (БФ + БЗ) + Решатель задач + Интерфейс.

Из этого определения понятно, что для каждого исследования создается своя интеллектуальная система типа ДСМ.

2. Почему ИС-ДСМ интеллектуальная система, а ДСМ-метод является методом интеллектуального анализа данных

Используя работу В.К. Финна [12], рассмотрим цепочку терминов: интеллектуальная способность (черта интеллекта) – интеллектуальная деятельность – интеллектуальная система типа ДСМ – интеллектуальный анализ данных.

2.1. Интеллектуальные способности (черты) интеллекта

(1) выделение существенного в данных;

(2) порождение последовательности «цель – план – действие» (целеполагание);

(3) отбор знаний (посылок для выводов, релевантных цели рассуждения);

(4) способность к рассуждению: вывод следствий из посылок, извлечение следствий посредством рассуждений, содержащих как правдоподобные выводы, используемые для выдвижения гипотез, так и достоверные (следовательно, под рассуждением понимаются последовательности правдоподобных и достоверных выводов);

(5) синтез познавательных процедур, реализующих амплиативные выводы (Ч.С. Пирс) – строение рассуждений (например, индукция + аналогия + абдукция → дедукция);

(6) рефлексия – оценка знаний и действий;

(7) способность к объяснению (ответ на вопрос «почему?»);

(8) аргументация при принятии решений;

(9) познавательное любопытство и способность к распознаванию (ответ на вопрос «что такое?»);

(10) способность к обучению и использование памяти;

(11) способность к уточнению неясных идей – преобразование их в понятия;

(12) способность к интеграции знаний для образования концепций и теорий;

(13) способность к изменению системы знаний при получении новых знаний и изменений ситуаций.

Интеллектуальная деятельность есть реализация интеллектуальных способностей, а интеллектуальная система – программная система, реализующая интеллектуальную деятельность.

В ИС-ДСМ интеллектуальные способности реализуются в двух режимах: автоматическом (способности 1, 3–10) и интерактивном (способности 2, 6, 11–13). Таким образом, ИС-ДСМ реализует интеллектуальную деятельность и, следовательно, является интеллектуальной.

Интеллектуальная система типа ДСМ реализует анализ данных, который будем называть интеллектуальным анализом данных.

2.2. Экспертная система и интеллектуальная система типа ДСМ: свойства

Экспертная система – одно из первых достижений искусственного интеллекта, которое начало использоваться и в медицине.

В базе знаний экспертной системы хранятся правила, связывающие признаки некоторого явления с его диагнозом. В этих правилах отражена интуиция эксперта, работающего в некоторой конкретной области.

Таблица 2

Свойства экспертной и интеллектуальной системы типа ДСМ

Свойства экспертной системы	Свойства ИС-ДСМ
Использование для решения задачи поверхностных, эвристических знаний, приобретенных от экспертов. Знания представляются в виде правил «если, то»	Обнаружение новых причинных знаний в данных (Knowledge Discovery) с помощью ДСМ-рассуждения или ДСМ-исследования
Имитация интуиции и опыта эксперта	Имитация и усиление познавательной деятельности
Машина вывода реализует вывод на знаниях	Решатель реализует ДСМ-рассуждение или ДСМ-исследование
Решение конкретной задачи	Осуществляется настройка универсального Решателя на конкретную задачу
Замкнутость теории, основанной на правилах, приобретенных от экспертов	Открытость порождаемой в результате анализа теории
Нет машинного обучения	Машинное обучение входит в процедуру индукции
Нет порождения новых знаний	Порождаются новые знания

Для приобретения знаний от экспертов инженеры по знаниям (когнитологи) и психологи создавали специальные методы их вербализации.

В 80-е годы прошлого века успех экспертных систем был столь грандиозным, что Э. Фейгенбаум – создатель экспертной системы *Dendral* – заявил: «Мы обнаружили, что лучше быть многознающим, чем умным». В дальнейшем Э. Фейгенбаум признал свою ошибку и создал систему *Meta-Dendral*, в которой было машинное обучение.

В табл. 2 приводятся свойства экспертной и интеллектуальной системы типа ДСМ.

III. ЭТАПЫ ПРОВЕДЕНИЯ ИССЛЕДОВАНИЯ В ПСИХИАТРИИ С ПОМОЩЬЮ ИС-ДСМ: ОБЩИЕ ЗАМЕЧАНИЯ ПЕРЕД РАССМОТРЕНИЕМ ПРИМЕРОВ

Самыми важными требованиями к исследованиям со стороны ДСМ-метода являются описание объектов дифференциальными признаками, использование положительных и отрицательных примеров исследуемого эффекта, описанных на одном и том же языке, т. е. с помощью одних и тех же признаков.

Первый этап: формирование содержательного представления проблемы, идеи, эмпирической теории, разработка языка описания объектов (создание совокупности дифференциальных признаков с их значениями), сбор данных и создание базы данных, из которой в дальнейшем будут формироваться базы положительных и отрицательных фактов для экспериментов на ИС-ДСМ. (Заметим, что в рамках одного исследования может быть проведено несколько экспериментов).

Создание языка описания данных в базе. Признаки описания психических расстройств и психосоциальных аспектов чаще всего берутся из шкал (примеры 2 и 3 из раздела IV настоящей работы: ДСМ-метод в психиатрии), разрабатываемых для исследования анкет (пример 4 из раздела IV). Части языка, относящиеся к заданию объективных данных, чаще всего принимаются такими, как это принято в предметной области, например, описания генов в примере 3 из раздела IV. Язык может быть разработан специально для исследования, как это было в исследовании психосоматического зуда (пример 1 из раздела IV). Сформулируем здесь некоторый принцип создания языка представления данных. Язык описания должен быть широким: специфика феномена, признаки, его определяющие (гипотезы), должны быть выявлены, порождены в результате эксперимента, как, например, признаки зуда, характерные именно для состояния «психосоматический зуд», выделенные из признаков описания зуда, который присутствует при многих заболеваниях. И еще одно требование к языку описания данных со стороны ДСМ-метода: язык должен быть един для описания положительных и отрицательных примеров (фактов), т.е. все они должны быть описаны одними и теми же признаками.

Значения признаков. ДСМ-метод реализует качественный анализ данных, это связано с операцией сходства, которая в нем используется. Поэтому для экспериментов на ИС-ДСМ осуществляется переход от числовых (количественных) значений признаков к категориальным (качественным, номинальным). Балльные оценки в шкалах и анкетах создаются для статистического анализа данных. Для использования ИС-ДСМ значения, применяемые в шкалах и анкетах, перед анализом либо заменяются на качественные – «сильно», «слабо», «да/нет», – либо сохраняются, но при этом рассматриваются как «символы» и «знаки», а не числа. Однако чаще всего используется значение «да/нет» (бинарный признак) для уменьшения перебора – сокращения времени экспериментов.

Положительные и отрицательные примеры. Врачи и эксперты используют «отрицательную» информацию в своей практике. Мышление медика, направленное на распознавание болезней, также реализует две принципиальные логические процедуры: «включение»/«исключение». Первая, собственно диагностика, состоит в обнаружении симптомов болезни, закономерно складывающихся в синдромы – проявления той или иной болезни (нозологии). Результатом (в идеале) является умозаключение об объяснимости конкретного заболевания (нозологии, предполагающей единство этио-

логии, патогенеза, клинических проявлений, динамики состояния, исходов) у конкретного пациента – *предварительный диагноз*. Вторая процедура «естественного» мышления врача, дифференциальная диагностика – исключение симптомов и синдромов, свойственных другим заболеваниям, которые могут иметь схожие проявления с конкретной клинической картиной, послужившей основой для предварительного диагноза. В результате составляется умозаключение о необъяснимости набора иных заболеваний у конкретного пациента, что подтверждает предварительный диагноз и превращает его в *диагноз клинический*.

Кроме того, анализ данных статистическими методами предполагает верификацию результатов анализа с помощью контрольной выборки, чаще всего в нее входят описания здоровых (с точки зрения цели анализа) людей.

Результат первого этапа – база данных, созданная с помощью табличного редактора (чаще всего с помощью Excel).

Второй этап: создание эксперимента на ИС-ДСМ, т. е. оформление содержания исследования и его проведение в парадигме ДСМ-метода.

Прежде всего, определяется, задается «эффект». Именно его наличие или отсутствие у объектов определяет отнесение этого объекта к базе положительных или отрицательных фактов. В принципе каждый из признаков может быть задан как эффект. Иногда рассматривается конъюнкция (или дизъюнкция) эффектов. Далее формируются базы положительных и отрицательных фактов (положительных и отрицательных примеров эффекта). Дополнительно создается база объектов, у которых нет данных о наличии или отсутствии рассматриваемого эффекта (τ-примеры), используемая для предсказаний с целью верификации результатов работы ИС-ДСМ.

Затем выполняются следующие этапы эксперимента:

1) применение ДСМ-рассуждения или ДСМ-исследования;

2) верификация результатов работы ИС-ДСМ (сравнение с оценками экспертов);

3) интерпретация результатов работы ИС-ДСМ;

4) внесение необходимых изменений и дополнений в представление знаний, использование других стратегий ДСМ-анализа, изменение используемых процедур Решателя ИС-ДСМ.

В психиатрии одно исследование может содержать несколько экспериментов. Это связано со следующими обстоятельствами: в рамках одного исследования могут рассматриваться несколько задач (как в примере 3 раздела IV), добавляться и удаляться признаки, изменяться значения признаков, использоваться разные методы и стратегии из арсенала ДСМ-подхода. И, наконец, могут проводиться эксперименты с использованием как ДСМ-рассуждения, так и с использованием ДСМ-исследований (обнаружение закономерностей в данных).

Для обозначения неоднократного экспериментирования в рамках одного исследования введем понятие «протокол эксперимента», в который входят:

- эффект (эффекты);
- база положительных фактов (примеров) эффекта;
- база отрицательных фактов (примеров) эффекта (отсутствия эффекта);
- названия методов, стратегий и т.д., которые использовались в эксперименте;
- результаты анализа;
- комментарии.

Протокол помогает структурировать проведение эксперимента.

IV. ДСМ-МЕТОД В ПСИХИАТРИИ: ПРИМЕРЫ ИССЛЕДОВАНИЙ С ПРЕДСТАВЛЕНИЕМ ДАННЫХ ДЛЯ АНАЛИЗА

Все эксперименты по исследованиям в области психиатрии проводились вместе с О.П. Шестерниковой с применением созданного ею Универсального решателя.

Пример 1. Исследование психосоматического (функционального) зуда.

В психосоматической медицине исследуются психические расстройства, проявляющиеся «телесными», или соматическими симптомами, напоминающими симптомы непсихических заболеваний. Опишем исследование, связанное с психосоматическим (функциональным) кожным зудом, т. е. с нарушениями кожной перцепции при наличии психического расстройства.

Психосоматический или функциональный зуд – аналог функциональной боли (например, «фантомной боли»), т. е. зуд без объективных кожных изменений (высыпаний). Дерматологическим «аналогом» такого зуда является зуд пруритогенный, т. е. возникающий в результате зудящего дерматоза (истинного кожного заболевания: атопический дерматит, экзема и т.п.), имеющий сугубо биологическую основу (например, действие медиаторов воспаления на периферические рецепторы восприятия зуда – прурицепторы – в коже).

У этого исследования две цели.

Первая – это диагностика, т.е. выявление симптомов зуда, как проявления психического расстройства, а также симптомов самого психического расстройства, т.е. выявление совокупности признаков, определяющих заболевание «психосоматический зуд» на фоне близких к нему по симптомам дерматологических заболеваний. Дело в том, что диагностика психосоматического зуда с помощью изучения описаний больных и интуиции медиков затруднена: и в случае функционального зуда высыпания тоже иногда присутствуют. Кроме того, возможна еще более сложная ситуация – имплицированный (амплифицированный) зуд, возникающий в тех случаях, когда имеющееся психическое расстройство сопровождается функциональным зудом, но возникает у пациента с «незудящим» кожным заболеванием (например, при угревой болезни, невусах и пр.) [13].

Вторая – это определение личностных и социальных признаков пациентов с психосоматическим зудом.

Для настоящего исследования в базе данных были представлены описания двух групп больных: в первую группу вошли больные, которым был поставлен диагноз психосоматический, или функциональный

зуд; вторая группа – это больные с имплицированным зудом: у них имеется психическое расстройство, кожный зуд и дерматологическое заболевание («незудящий дерматоз»), в виде высыпаний на коже.

В язык описания пациентов вошли признаки психических расстройств, зуда и дерматологических заболеваний, при этом были введены признаки, достаточно широко описывающие эти расстройства. Для поисковых исследований, проводимых с помощью ИС-ДСМ, язык описания данных должен быть шире, чем априорные предположения экспертов. Сочетания наборов признаков, подтверждающих или опровергающих предположения экспертов, должны быть порождены в ходе экспериментов с помощью ИС-ДСМ. В данном исследовании должны быть порождены сочетания признаков психических расстройств и зуда, характерные именно для заболевания психосоматический, функциональный зуд.

В язык описания данных вошли следующие признаки:

признаки психиатрических расстройств – были ли у пациента любые психические расстройства на протяжении жизни (в анамнезе), лечение и наблюдение у психиатра в анамнезе, степень тяжести психического расстройства в анамнезе, имеется ли психическое расстройство на момент обследования, нозологический диагноз психического заболевания на момент обследования, есть ли депрессия в анамнезе, степень тяжести депрессии в анамнезе, тип течения депрессии, есть ли у пациента признаки тревоги, мысли о суициде, мысли о суициде из-за заболевания кожи, частота посещения мыслей о суициде из-за заболевания кожи, расстройства тревожного спектра, сезонное улучшение или ухудшение психического состояния, суточное улучшение или ухудшение психического состояния;

описание кожного заболевания – длительность заболевания кожи, длительность симптомов, частота развития обострений кожных заболеваний, длительность зуда, частота возникновения зуда, зуд в настоящее время, зуд в течение прошлого года, время обострения зуда по суточному ритму, время обострения зуда по сезонному ритму;

характеристики зуда и особенности кожных заболеваний – локализация кожных высыпаний, сопутствующие зуду симптомы, расположение на туловище, расположение на конечностях, дерматологический диагноз и тяжесть, качественные характеристики зуда, интенсивность зуда по визуальной аналоговой шкале, характеристики зуда по Эппендорфскому опроснику зуда, результаты расчесов;

хронические заболевания – кардиопаталогия, респираторные заболевания, диабет, ревматология;

социодемографические признаки – возраст, пол, брачный статус; образование, социальный статус; экономические трудности в течение последних 5 лет, стрессовые жизненные события в течение последних 6 месяцев, зависимость от психоактивных веществ или алкоголя, самооценка здоровья.

Эксперимент. В базу положительных фактов (положительные примеры) вошли больные, которым был поставлен диагноз психосоматический, или функциональный зуд: у них есть психическое расстройство (шизофрения, аффективное заболевание, невротическая патология), кожный зуд, но отсутствуют объективно диагностируемые кожные высыпания.

Другая группа (отрицательные примеры) – это больные с имплицированным зудом: у них есть психическое расстройство, дерматологическое заболевание («незудящий дерматоз»), проявляющееся в высыпаниях на коже, и кожный зуд. Фактически, с точки зрения дерматолога, такие пациенты имеют на коже первичные высыпания, которые чесаться «не должны», но при этом они высказывают жалобы на зуд (иногда достаточно интенсивный).

В результате работы ИС-ДСМ были порождены сочетания признаков психических расстройств и признаков зуда, характерных для самостоятельного психосоматического расстройства [14]:

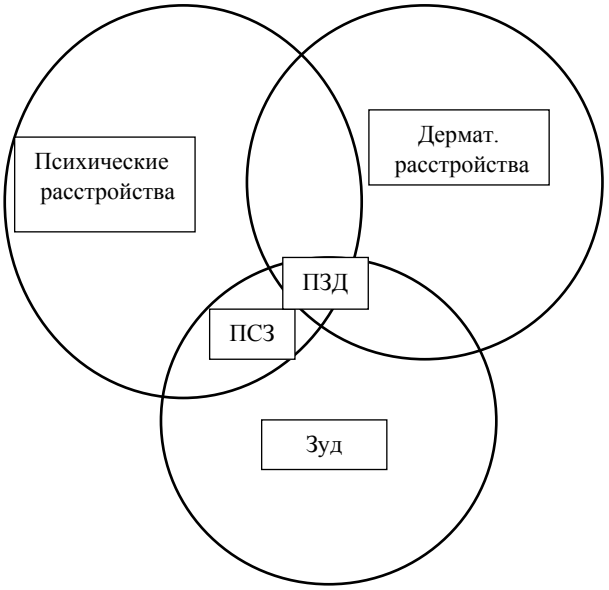
- интенсивность кожного зуда (часто значительная) не соответствует выраженности (часто минимальной) объективно наблюдаемых высыпаний;
- охват участков кожи, значительно выходящих за пределы имеющихся элементов сыпи (зуд «чистой кожи»);
- персистирование жалоб на зуд, длительно сохраняющихся после редукции высыпаний в периоде объективно регистрируемой ремиссии дерматоза;
- эпизоды зуда, не сопровождающегося специфическими высыпаниями вне обострений кожного заболевания;
- своеобразие описаний кожных ощущений, нетипичных для зудящих дерматозов («ужаливание», «ползанье», «покусывание», «мурашки»);
- отсутствие выраженной аутодеструкции (расчесов), соотносимое с поведением при кожном органном неврозе (соматоформном зуде *sine materia*) – щадящим обращением с кожей.

Психосоматический зуд представляет собой расстройство системы «психика – кожа» [15].

Для уточнения признаков психосоматического зуда можно было провести еще три эксперимента: в первом в качестве отрицательных примеров отобрать из базы данных описания пациентов только с зудом, во втором – только с психическими расстройствами, в третьем – только с дерматологическими расстройствами. На рисунке изображены области положительных и отрицательных примеров.

Пример 2. Диагностика и исследование депрессии с использованием шкал Гамильтона и *PHQ*

В первом варианте исследования депрессии было предложено использование для описания пациентов (объектов) признаков, содержащихся в шкале Гамильтона (см. табл. 1). Предполагалось в базу положительных фактов отнести описания пациентов с оценкой тяжести депрессии с суммой баллов ≥ 7, в базу отрицательных фактов – <7 (норма). Однако эксперты отвергли это предложение, так как шкала Гамильтона используется для объективизации тяжести клинически верифицированной депрессии у пациента, а «норма» используется для оценки состояния пациента после лечения.



Области положительных и отрицательных примеров
(ПСЗ – психосоматический зуд, ПЗД – Психиатрическое расстройство + Зуд + Дерматологическое расстройство).

Таблица 3

Шкала *Patient Health Questionnaire (PHQ)*

№	Симптом	Оценка в баллах*
1	Отсутствие интереса к происходящим событиям	От 0 до 3
2	Безразличие, подавленность	От 0 до 3
3	Проблемы с засыпанием, бессонница, наоборот, слишком много спали	От 0 до 3
4	Чувство усталости или упадок сил	От 0 до 3
5	Отсутствие аппетита или переедание	От 0 до 3
6	Трудно сосредоточиться на чтение или просмотре телевизора	От до 3
7	Двигаетесь или говорите необыкновенно медленно (заторможенно) или наоборот, двигаетесь больше, чем обычно	От 0 до 3
8	Мысли о самоубийстве или о причинении себе вреда	От 0 до 3

* Оценки: 0 – не каждый день; 1 – несколько дней; 2 – более, чем в половине дней; 3 – почти каждый день.

В результате был создан второй вариант исследования депрессии: для этого в базу данных вошли пациенты, обследованные с помощью шкалы Гамильтона, а также здоровые люди, обследованные с помощью шкалы *PHQ* (*Patient Health Questionnaire*) (табл. 3). Она была разработана доктором медицинских наук Робертом Л. Спитцером и его коллегами из Pfizer Inc и показала свою эффективность на практике. Шкала *PHQ* служит для описания некоторых депрессивных признаков у здоровых людей (для скрининга на депрессию здоровых людей).

После заполнения шкалы баллы суммируются. Полученные результаты интерпретируются:

менее 4 – минимальная депрессия;
5-9 – легкая депрессия;
10-14 – умеренная депрессия;
15-16 – тяжелая депрессия;
более 20 – крайне тяжелая депрессия.

В качестве языка описания объектов из созданной для исследования базы используется объединение признаков, содержащихся в шкалах Гамильтона и *PHQ*.

Эксперимент. В базу положительных фактов вошли описания пациентов с суммарной оценкой по шкале Гамильтона ≥ 18, а также «здоровых» людей с суммарной оценкой по шкале *PHQ* ≥ 15. В базу отрицательных фактов вошли все остальные пациенты и «здоровые» люди. Для эксперимента в реальное время (для сокращения перебора) значения признаков были преобразованы в значения «да» или «нет». Оценка «да» интерпретируется как сильная выраженность признака, оценка «нет» – как слабая выраженность признака. В шкале Гамильтона часть симптомов имеет значения от 0 до 4 баллов, часть – от 0 до 2 баллов. Преобразование осуществляется следующим образом. Для симптомов с оценкой от 0 до 4 баллов: значения в баллах 0 и 1 преобразуются в

значение «нет»; значения в баллах 2, 3, 4 – в значение «да». Для симптомов с оценкой от 0 до 2 баллов: значения в баллах 0 и 1 преобразуются в значение «нет», значение 2 балла – в значение «да». Для шкалы *PHQ*: значения 0, 1 балла преобразуются в значение «нет»; значения 2, 3 балла – в значение «да».

В результате работы ИС-ДСМ были порождены гипотезы, в которые входят индивидуальные первичные признаки, свернутые в комплексные характеристики, определяющие наличие депрессии и её отсутствие.

Сравнение результатов диагностики депрессии с помощью ИС-ДСМ и шкал Гамильтона и *PHQ* (с суммированием значений признаков депрессии) можно рассматривать как тест Тьюринга для ИС-ДСМ.

Полученные с помощью ИС-ДСМ гипотезы могут служить основанием для типологизации депрессии. Задача типологизации часто встречается в исследованиях социальных явлений с применением ДСМ-метода. В работе [16] социальные группы, полученные в результате типологизации, рассматриваются как метафоры, им даются специальные названия. В психиатрии метафорам соответствуют прототипы психических расстройств, которые есть в арсенале у каждого врача. Такое исследование с помощью ДСМ-метода можно рассматривать как объективизацию этих прототипов.

Можно поставить задачу дифференциации депрессии от другого психического расстройства. В этом случае язык описания данных должен включать признаки как депрессии, так и этого психического расстройства, так как ДСМ-метод требует описания положительных (депрессия) и отрицательных примеров (другое психическое расстройство) на едином языке.

Пример 3. Исследование связи некоторых генетических характеристик и эмоционального и/или волевого дефицита как проявления негативного синдрома у пациентов с расстройствами шизофренического спектра

Для оценки позитивного и негативного синдромов шизофрении психиатры во всем мире используют чаще всего шкалу *PANSS* (табл. 4), которая была разработана в 80-е гг. XX в. (С.П. Кей, Л.А. Оплер и А. Фицбейн).

Таблица 4

***PANSS*: шкала оценки позитивного и негативного синдромов шизофрении**

Симптомы	**Оценка тяжести симптома в баллах**
Шкала позитивного синдрома (P)	
P1. Бред	От 1 до 7
P2. Дезорганизация мышления	От 1 до 7
P3. Галлюцинаторное поведение	От 1 до 7
P4. Психомоторное возбуждение	От 1 до 7
P5. Грандиозность	От 1 до 7
P6. Подозрительность/преследование	От 1 до 7
P7. Враждебность	От 1 до 7
Шкала негативного синдрома (N)	
N1. Уплощенный аффект	От 1 до 7
N2. Эмоциональная отстраненность	От 1 до 7
N3. Снижение коммуникабельности	От 1 до 7
N4. Пассивная/апатическая социальная самоизоляция	От 1 до 7
N5. Нарушения абстрактного мышления	От 1 до 7
N6. Снижение спонтанности и речевой активности	От 1 до 7
N7. Стереотипность мышления	От 1 до 7
Шкала общего психопатологического синдрома (G)	
G1. Соматизация	От 1 до 7
G2. Тревога	От 1 до 7
G3. Чувство вины	От 1 до 7
G4. Напряженность	От 1 до 7
G5. Манерность и поза	От 1 до 7
G6. Депрессия	От 1 до 7
G7. Двигательная заторможенность	От 1 до 7
G8. Некооперативность	От 1 до 7
G9. Необычное содержание мышления	От 1 до 7
G10. Дезориентировка	От 1 до 7
G11. Нарушения внимания	От 1 до 7
G12. Нарушение суждений и критики	От 1 до 7
G13. Волевые нарушения	От 1 до 7
G14. Снижение контроля побуждений	От 1 до 7
G15. Аутизация	От 1 до 7
G16. Активная социальная изоляция	От 1 до 7

Методика использования шкалы *PANSS* следующая: значения тяжести симптомов складываются отдельно для позитивного и негативного синдромов. Оценка выраженности позитивного или негативного синдрома определяется по большему значению из полученных сумм. Такая классификация (на позитивный и негативный синдромы) важна для назначения лечения.

Цель исследования: обнаружить у пациентов с расстройством шизофренического спектра при установленном негативном синдроме связь некоторых генетических характеристик с отсутствием или со степенью тяжести эмоционального и/или волевого дефицита.

Наряду с шизофренией к патологии шизофренического спектра относятся следующие психические расстройства:

- шизоаффективное расстройство,
- шизофрениформное расстройство,
- бредовое расстройство,
- шизоидное расстройство личности,
- шизотипическое расстройство.

Итак, исследуем связь генов BDNF, STR2A, HTTLPR и ZNF804A с эмоциональным/волевым снижением у пациентов с расстройством шизофренического спектра. Именно с этими генами эксперты предполагают связь эмоционального/волевого снижения у пациентов. (Об исследовании ассоциации вариации (rs1344706) в гене ЗНФ804А с шизофренией и ее симптомами с помощью статистических методов см. в [17]).

Признаки эмоционального снижения: N1, N2, N3, N4 и N6 из негативного синдрома, признаки волевого снижения: G13 и G 16 из общего патологического синдрома (см. шкалу *PANSS*).

Эксперимент 1. Установление связи генов с наличием и отсутствием эмоционального/волевого снижения у пациентов с расстройством шизофренического спектра.

Для уменьшения перебора при анализе с помощью ИС-ДСМ был выполнен переход от балльной оценки признаков к бинарной оценке («да/нет»): 1 балл («нет симптома») – «нет»; 2, 3, 4, 5, 6 баллов («симптом слабо выражен», «симптом средне/сильно выражен») – «да».

Исследовались связи каждого из признаков N1, N2, N3, N4, N6, G13, G 16, их пар, троек и т.д. со всеми предоставленными для анализа генами. При этом определялось сходство по аллелям всех трех генотипов, входящих в состав каждого гена.

Эксперимент 2. Установление связи генов с тяжестью эмоционального/волевого снижения у пациентов с расстройством шизофренического спектра.

В этом эксперименте был выполнен переход от балльной оценки признаков к бинарной оценке («да/нет»): 2, 3 балла («симптом слабо выражен») – «нет»; 4, 5, 6 баллов («симптом средне/сильно выражен») – «да».

Заметим, что если эксперты добавят соответствующие генетические признаки, то у пациентов можно установить связь этих признаков с другими симптомами шкалы *PANSS*.

Пример 4. Исследование связи депрессии, тревоги, стигматизации и расстройства образа тела (дисморфического расстройства) с наличием кожных заболеваний, а также связей этих психических расстройств между собой.

Данные для исследования получены из протокола исследования Международного проекта «Психосоциальное бремя болезней кожи» Европейского общества дерматологов и психиатров (ESDaP) (16 европейских стран) [18].

Социальная стигматизация определяется как опыт социального неодобрения, дискредитации или девальвации, основанной на качественных признаках или оценке окружающими физических характеристик субъекта. Нарушения образа тела – это состояния, связанные с недовольством внешним обликом собственного тела, в том числе видом кожи. Для некоторых людей степень неудовлетворенности настолько высока, что это состояние проявляется в психическом расстройстве и обозначается как дисморфическое расстройство. Часто именно кожные заболевания приводят к возникновению этих психических расстройств.

Источниками исходной информации были самоопросники (шкала образа тела – *Dysmorphic Concern Questionnaire* (*DCQ*) и шкала стигматизации – *Perceived Stigmatization Questionnaire* (*PSQ*)). На их основе создан язык описания данных для анализа с помощью ИС-ДСМ. Для сокращения времени проведения экспериментов был выполнен перевод балльных оценок состояния, содержащихся в ответах на вопросы, в значения «да/нет» (бинарный категориальный признак).

Исследование состоит из нескольких экспериментов. В каждом из них в базу положительных фактов входят описания пациентов с одним, двумя, тремя или четырьмя исследуемыми психическими расстройствами и с наличием кожных заболеваний. В базу отрицательных фактов входят описания тех же психических расстройств, но без кожных заболеваний.

Постановки других задач и представление данных для их решения с помощью ИС-ДСМ содержатся в [19].

ЗАКЛЮЧЕНИЕ

На основе нескольких примеров описана организация исследований в области психиатрии и подготовки данных для их проведения с помощью интеллектуальной системы типа ДСМ (ИС-ДСМ), реализующей интеллектуальный анализ данных на основе ДСМ-метода.

Сформулируем несколько принципов подготовки данных для проведения исследований с применением ИС-ДСМ.

Важный этап подготовки данных – разработка языка описания объектов, создание совокупности дифференциальных признаков с их значениями.

1. Положительные и отрицательные примеры (факты) эффекта должны быть описаны на едином языке, т.е. содержать одни и те же признаки.

2. Язык описания данных должен быть широким: специфика феномена, признаки, его определяющие (гипотезы), могут быть выявлены, т. е. порождены в результате эксперимента.

3. Для использования ИС-ДСМ значения, применяемые в шкалах и анкетах, перед анализом либо заменяются на качественные – «сильно», «слабо», «да/нет», – либо сохраняются, но при этом рассматриваются как «символы» и «знаки», а не числа. Однако чаще всего используются значения «да/нет» (бинарный признак) для уменьшения перебора – сокращения времени экспериментов.

В психиатрии одно исследование может содержать несколько экспериментов. Это связано с тем. что в рамках одного исследования могут рассматриваться несколько задач (как в примере 3 раздела IV), возможно добавление и удаление признаков, изменение значений признаков, использование различных методов и стратегий из арсенала ДСМ-подхода. И, наконец, могут проводиться эксперименты с использованием как ДСМ-рассуждения, так и ДСМ-исследований (обнаружение закономерностей в данных).

4. Для отображения неоднократного экспериментирования в рамках одного исследования вводится понятие «протокол эксперимента», в котором следует отразить:

- эффект (эффекты);
- базу положительных фактов (примеров) эффекта;
- базу отрицательных фактов (примеров) эффекта (отсутствия эффекта);
- названия методов, стратегий и т.д., которые использовались в эксперименте;
- результаты анализа;
- комментарии.

Протокол помогает структурировать проведение эксперимента.

Настоящая работа может служить источником методических сведений по подготовке данных для их анализа с помощью ИС-ДСМ.

СПИСОК ЛИТЕРАТУРЫ

1. DSM-5 / Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, American Psychiatric Association, 2013. – 991 p.
2. Павличенко А.В. Настоящее и будущее диагноза в психиатрической практике // Трудный пациент. – 2015. – Т.13, № 5-6. – С. 41-49.
3. Виктор Константинович Финн: библиогр. указ. / Минобрнауки России, Фед. гос. бюд. образоват. учреждение высш. образования «Рос. гос. гуманитарный ун-т» (РГГУ); [сост.: М.А. Михеенкова, А.В. Кученкова; под общ. ред. Л.Н. Простоволосовой; вступ. ст. Д.А. Поспелов, Д.Г. Лахути, В.Б. Тарасов]. – 2-е изд., доп. – Москва: РГГУ, 2018. – 117 с. – (Ученые РГГУ).
4. Шестерникова О.П., Агафонов М.А., Винокурова Л.В., Панкратова Е.С., Финн В.К. Интеллектуальная система прогнозирования развития сахарного диабета у больных хроническим панкреатитом // Искусственный интеллект и принятие решений. –2015. – № 4. – С. 12-50.
5. Шестерникова О.П., Финн В.К., Винокурова Л.В., Лесько К.А., Варванина Г.Г., Тюляева Е.Ю. Интеллектуальная система для диагностики заболеваний поджелудочной железы // Научно-техническая информация. Сер. 2. – 2019. – № 10. – С. 41 – 48.
6. Классы МКБ-10 / F00 – F99 Психические расстройства и расстройства поведения. – URL: http://mkb-10.com.
7. Психометрические шкалы. – URL: http://ncpz.ru.
8. Автоматическое порождение гипотез в интеллектуальных системах / под общ. ред. В.К. Финна. – Москва: Изд-во URSS, 2020. – 526 с.
9. Финн В.К. Эвристика обнаружения эмпирических закономерностей и принципы интеллектуального анализа данных // Искусственный интеллект и принятие решений. – 2018. – № 3. – С. 3-19.
10. Финн В.К. Шестерникова О.П. Эвристика обнаружения эмпирических закономерностей посредством ДСМ-рассуждений // Научно-техническая информация. Сер. 2. – 2018. – № 9. – С. 7-42.
11. Финн В.К. Словарь терминов искусственного интеллекта // В кн. Интеллект, информационное общество, гуманитарное знание и образование. – Москва: Изд-во URSS, 2021. – С. 437-438.
12. Финн В.К. К структурной когнитологии: феноменология сознания с точки зрения искусственного интеллекта // В кн.: Искусственный интеллект: методология, применения, философия. – М. – КРАСАНД, 2018. – С. 256-277.
13. Львов А.Н., Бобко С.И., Романов Д.В. Соматоформный и амплифицированный зуд // Российский журнал кожных и венерических болезней. – 2013. – № 4. – С. 39-43.
14. Романов Д.В., Финн В.К., Фабрикантова Е.Ф., Андрющенко А.В., Львов А.Н., Бобко С.И. Интеллектуальная система типа ДСМ для автоматизированной поддержки исследования коэнестезиопатических расстройств (зуда) в дерматологической практике: методология и некоторые результаты // Психические расстройства в общей медицине. – 2015. – № 4. – С. 30-39.
15. Фабрикантова Е.Ф. Моделирование систем средствами искусственного интеллекта // Научно-техническая информация. Сер 2. – 2012. – № 12. – С. 1-7.
16. Михеенкова М.А., Климова С.Г. Интеллектуальный анализ данных в социологических исследованиях // Научно-техническая информация. Сер. 2. – 2018. – № 12. – С. 12–24.
17. Lezheiko T.V., Romanov D.V., Kolesina N.Yu., Golimbet V.E. Data on association of the variation (rs1344706) in the ZNF804A gene with schizophrenia and its symptoms in the Russian population // Data in brief 24 (2019) 103985 (Исследование ассоциации вариации (rs1344706) в гене ЗНФ804А с шизофренией и ее симптомами у населения России).
18. Dalgard F.J., Bewley A., Evers A.W., Gieler U., Lien L., Sampogna F., Ständer S., Tomas-Aragones L., Vulink N., Kupfer J. Stigmatisation and body image impairment in dermatological patients: protocol for an observational multicentre study in 16 European countries. BMJ Open. 2018 Dec 22;8(12): e024877. DOI: 10.1136/bmjopen-2018-024877. PMID: 30580274; PMCID: PMC6307615.
19. Фабрикантова Е.Ф. Применение ДСМ-метода для исследования расстройств шизофренического спектра // Труды шестнадцатой национальной

конференции по искусственному интеллекту с международным участием КИИ-2018 (Москва, 2018 г.). – М: РКП, 2018. Том 1. – С. 239-244. – URL: https://elibrary.ru/item.asp?id=35568493.

Материал поступил в редакцию 08.02.21.

ФАБРИКАНТОВА Елена Федоровна – кандидат технических наук, старший научный сотрудник Федерального исследовательского центра «Информатика и Управление» (ФИЦ ИУ) РАН, Москва
E-mail: el.fabrikantova@yandex.ru).

РОМАНОВ Дмитрий Владимирович – доктор медицинских наук, профессор кафедры психиатрии и психосоматики ИКМ ФГАОУ ВО Первый МГМУ им. И.М. Сеченова (Сеченовский Университет), ведущий научный сотрудник отдела по изучению пограничной психической патологии и психосоматических расстройств ФГБНУ Научный центр психического здоровья; врач-психиатр ГБУЗ Научно-практический центр дермато-венерологии и косметологии ДЗ г. Москвы; ORCID: 0000-0002-1822-8973; РИНЦ AuthorID: 189414.
E-mail: newt777@mail.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322’366

Ал-др А. Хорошилов, Ю.В. Никитин, С.И. Пшеничный, М.А. Шевкунов, А.А. Хорошилов

Морфологический анализатор МетаФраз нового поколения

Описана концепция, предложены методы и алгоритмы создания морфологического анализатора по технологиям МетаФраз нового поколения, в основу которого, так же как и анализатора МетаФраз первого поколения, положена языковая морфологическая модель флективных классов русских слов, предложенная профессором Г.Г. Белоноговым в конце 60-х гг. прошлого столетия. В концепции содержится ряд положений, обеспечивающих высокое быстродействие и качество обработки текстовых форм слов. Оптимальное соотношение различных типов словарей в составе декларативных средств анализатора, а также применение быстродействующих процедур поиска в этих словарях могут обеспечить требуемые характеристики нового поколения анализатора. В соответствии с предложенными проектными решениями разработан макет морфологического анализатора МетаФраз второго поколения, опытная эксплуатация которого показала его работоспособность и возможность достижения им требуемых технологических и эксплуатационных характеристик.

Ключевые слова: *флективные классы русских слов, морфологический анализатор, машинная грамматика, морфологический анализ, словоизменение, словообразование, программные средства, декларативные средства, автоматическая нормализация слов*

DOI: 10.36535/0548-0027-2021-04-3

ВВЕДЕНИЕ

Стремительный рост объемов текстовой информации в сети Интернет и необходимость обеспечения доступа к ней значительно ускорили развитие технологий машинной обработки текстов на естественном языке (*Natural Language Processing – NLP*) [1]. Эти технологии ориентированы на преобразование текста в его формализованное представление в виде дискретных или комбинаторных структур для дальнейшего выполнения различных аналитических операций. Такое преобразование возможно выполнить с помощью многоступенчатого комплекса процедур морфологического, семантико-синтаксического и концептуального анализа и синтеза текстов. На каждом этапе этого комплекса производится формальное преобразование иерархии единиц смысла текста, выраженных словами, словосочетаниями, предложениями и содержанием текстов. Цель подобного преобразования для каждой единицы смысла – построение формальной модели. Так, например, на этапе обработки слова осуществляется морфологический анализ, обеспечивающий построение его формальной модели на основе информации о буквенном коде. Важной характеристикой процедуры построения морфологической модели является ее быстродействие и точность назначения грамматических характеристик. Эта процедура, как правило, предшествует всем остальными процедурам автоматической обработки текста, и именно она оказывает наибольшее воздействие на процедуры обработки более крупных фрагментов текста.

МАШИННАЯ ГРАММАТИКА НА ОСНОВЕ ФЛЕКТИВНЫХ КЛАССОВ

Под термином «машинная грамматика» понимается комплекс формальных правил, процедур и декларативных средств, обеспечивающих автоматическое преобразование текстовых представлений слов в формальное описание модели в виде совокупности грамматических и семантических характеристик [2]. Обычно машинные грамматики базируются на общепринятых грамматических правилах и формализмах, полученных в результате выявленных закономерностей функционирования языка и речи. Таким формализмом в наших исследованиях является разработанная профессором Г.Г. Белоноговым *система флективных классов слов русского языка* [2, 3]. Её можно рассматривать как универсальную классификацию русских слов, в которой представлены их основные типы словоизменительных парадигм.

Эта классификация положена в основу разработанной нами морфологической модели для русского языка, в рамках которой на основе методов лингвистической аналогии были установлены закономерности между грамматическими характеристиками слов и их конечным буквенным составом, а также предложены методы формального деления слов на слова с регулярной или с аномальной системой словоизменения и словообразования. Такая формальная модель обеспечила возможность автоматического назначения словам грамматических характеристик по их буквенному коду [4].

Система грамматических характеристик классов слов содержится в декларативных средствах, необходимых для функционирования процедур морфологического анализа. Создание декларативных средств больших объемов обеспечивается возможностью как автоматического разделения слов на различные категории, так и автоматического контролируемого назначения им специфических наборов грамматических и семантических характеристик в рамках используемой формальной языковой модели.

В основе концепции и алгоритмических решений морфоанализатора МетаФраз нового (второго) поколения лежит та же языковая модель, что была использована в морфоанализаторе МетаФраз первого поколения [2]. Она базируется на морфологической модели, основанной на системе флективных классов слов и принципе лингвистической аналогии, заложенных в ее реализацию. Принцип лингвистической аналогии опирается на гипотезу, утверждающую, что *в русском языке объективно существует сильная корреляция между конечным буквосочетанием слов и их грамматическими характеристиками.* Для реализации этого принципа разработчиками был создан эталонный словарь конечных буквосочетаний словоформ, представляющий все возможные конечные буквосочетания форм слов в различных контекстных окружениях [5].

При разработке морфологического анализатора МетаФраз первого поколения в качестве основной использовалась приближенная процедура анализа словоформ, выполняемая на основе метода аналогии. При осуществлении этой процедуры было установлено, что назначение грамматических характеристик на основе принципа аналогии возможно только для словоформ, имеющих регулярную трансформационную систему словоизменения. Наличие в словаре конечных буквосочетаний слов с нерегулярной (аномальной) системой словоизменения и словообразования приводило к существенным нарушениям при использовании этого принципа. Поэтому предварительно было необходимо исключить из словаря конечных буквосочетаний слова с аномальной трансформационной системой словоизменения, к которым были отнесены служебные слова, супплетивные формы слов и слова, длина которых не превышала пяти символов. Все они были помещены в словарь, названный словарем коротких и служебных слов.

Ориентация на словари словоформ слов обеспечила высокое быстродействие и точность обработки, поскольку здесь не было необходимости использовать достаточно ресурсоемкие вычислительные действия по установлению основы (или псевдоосновы) и нахождения совместимого грамматического окончания (или псевдоокончания).

Таким образом, в состав словарей морфоанализатора МетаФраз первого поколения первоначально включены только два словаря: словарь коротких и служебных слов и словарь конечных буквосочетаний. В более поздних версиях для большей экономии памяти дополнительно был разработан словарь (таблица) ФКГИ (флективный класс – грамматическая информация), обеспечивающий возможность преобразования базового набора грамматических характеристик[1] в их полный состав и позволяющий существенно сократить объем словарей коротких и служебных слов и конечных буквосочетаний [2].

Обработка слов при помощи алгоритма морфологического анализатора выполняется следующим образом: вначале производится поиск анализируемой формы слова в словаре коротких и служебных слов и, если она там находится, ей назначается базовый набор грамматических характеристик. Если эта словоформа не была обнаружено в словаре коротких и служебных слов, то для нее производится инверсия буквенного состава и выполняется поиск на наибольшее совпадение её конечного буквосочетания с буквосочетанием одного из элементов словаря конечных буквосочетаний. После нахождения подходящего буквосочетания анализируемой словоформе назначаются грамматические характеристики. Недостающая грамматическая информация устанавливается по таблице ФКГИ.

Несколько слов о технологиях создания и ведения этих словарей. Здесь нужно отметить, что относительно простая реализация анализатора, выполненная путем поиска в словаре коротких и служебных слов и словаре конечных буквосочетаний, обеспечивалась сложным комплексом процедур создания и ведения этих словарей. Так, например, если относительно просто выделить короткие и служебные слова, состав которых невелик, то отделить слова, имеющие аномальную систему словоизменительных трансформаций, было значительно сложнее. Это обеспечивалось процедурами анализа их буквенного состава и синтеза всех форм словоизменительных и словообразовательных форм слов, а также процедурами морфемного анализа и проверки совместимости между собой всех типов морфем, составляющих эти словоформы. Достаточно сложной являлась и процедура формирования словаря конечных буквосочетаний, создаваемых на основе больших корпусов текстов. Здесь наряду с задачей определения грамматических признаков было необходимо автоматически определять словоформы с регулярной трансформационной системой словоизменения и словоформы с аномальной трансформационной системой. Для словоформ с регулярной системой словоизменения определялись конечные

[1] К базовым характеристикам словоформы относятся номер флективного класса и грамматическое окончание. Эти характеристики могут обеспечить генерацию более полного набора грамматической информации формы слова: индекс грамматического класса, значений рода, числа, падежа, лица и одушевленности.

буквосочетания, однозначно определяющие грамматические характеристики анализируемой словоформы. Выявленные словоформы с аномальной трансформационной системой переносились в словарь коротких и служебных слов. Для реализации этих задач был разработан комплекс программ в составе Автоматизированной словарной службы.

Необходимо отметить, что концепция МетаФраз, изначально ориентированная на реализацию принципа лингвистической аналогии, была революционной, поскольку предполагалось, что основной поток слов будет обрабатываться не точной процедурой, как это принято в традиционных процедурах морфологического анализа, а приближенной процедурой, которая обычно применялась только для обработки «новых» слов. Точному анализу по словарю словоформ должно подвергаться относительно небольшое число слов, содержащихся в словаре коротких и служебных слов. Еще одним базовым принципом концепции была ориентация на словари словоформ и обработку конкретных форм слов. Был еще и третий принцип – максимальная экономия памяти ЭВМ[2], что объяснялось возможностями ранних этапов развития вычислительной техники. Принципом экономии памяти были пронизаны все применяемые технологические решения, иногда в ущерб качеству и быстродействию. Технология построения словаря конечных буквосочетаний обеспечивала максимальное его сжатие, как по составу элементов, так и по длине эталонных конечных буквосочетаний. Это в значительной степени определяет точность анализа слов на основе принципа лингвистической аналогии. В технологии *создания и ведения словарей* обязательно должна быть операция проверки на совместимость всех элементов морфемной структуры слов в рамках их словоизменительных парадигм.

ДЕКЛАРАТИВНЫЕ СРЕДСТВА МОРФОАНАЛИЗАТОРА МЕТАФРАЗ ВТОРОГО ПОКОЛЕНИЯ

Длительная эксплуатация анализатора МетаФраз первого поколения позволила выявить не только его значительные преимущества, выражающиеся в компактности, быстродействии и высокой точности назначения грамматических характеристик текстовым словам (свыше 95%), но и ряд существенных недостатков. К ним можно отнести относительно небольшой состав грамматических характеристик форм слов, назначаемых в процессе их обработки, а также невозможность автоматизированного назначения словообразовательных и семантических характеристик, присущих всем членам словоизменительной парадигмы слова; например, таких как характеристика глагольности существительных и прилагательных, признак одушевленности существительных, супплетивности прилагательных, а также ряд специфичных признаков принадлежности к группам слов, реализующих важные синтаксические функции при анализе структуры предложения.

Как было отмечено выше, изначальная ориентация на словари словоформ позволила существенно упросить алгоритмы анализа форм слов. Но нужно иметь в виду, что основы слов представляют всю совокупность форм словоизменительной парадигмы, а словари словоформ – только одну форму парадигмы. Поэтому может показаться, что для одинакового покрытия текстов требуется многократное увеличение объемов словарей словоформ. Но, как оказалось, это не совсем так. Наши исследования показали, что при одинаковом покрытии текстов увеличение объемов словарей словоформ может быть не столь значительным. Объемы словарей словоформ и словарей основ слов при одинаковом покрытии текстов находились в отношении 2,5:1, т. е. в среднем в текстах из каждой словоизменительной словоформы содержалось только по 2,5 формы слов из каждой их парадигмы.

Но для существенного повышения покрывающей способности словарей все же словари основ слов предпочтительнее, поскольку они позволяют хранить словообразующую и семантическую информацию для всех членов словоизменительной парадигмы. В связи с этим в концепции МетаФраз второго поколения должна быть предусмотрена возможность использования не только словарей словоформ, но и словарей основ слов без значительного увеличения вычислительной сложности алгоритмов их обработки. Для того, чтобы значительно повысить технологические и эксплуатационные характеристики морфоанализатора МетаФраз нового поколения, необходимо добиться:

1) повышения быстродействия функционирования анализатора не менее чем на 50%;

2) повышения качества декларативных средств анализатора (увеличение совокупности грамматических и семантических характеристик слов до 30 элементов) и значительного увеличения их покрывающей способности;

3) сокращения трудозатрат на разработку декларативных средств и программного обеспечения анализатора не менее чем на 50%.

Возможными путями реализации поставленных задач могут быть следующие решения:

- повышение быстродействия анализатора можно обеспечить включением в его состав дополнительной быстродействующей процедуры, обрабатывающей основной поток (не менее 80%) текстовых словоформ. Такая процедура должна использовать словарь, содержащий полный набор грамматических и семантических характеристик наиболее часто употребляемых форм слов русского языка;
- повышение качества декларативных средств возможно достичь при расширении спектра грамматических (формообразующих и словообразующих) и семантических характеристик. Автоматизированное назначение этих характеристик как словоформам, так и нормальным формам слов должно обеспечиваться контролируемыми технологическими процессами создания и ведения словарей. Формат словарных статей должен быть реализован в виде списковой структуры (название признака – значение признака);

[2] В ранних реализациях объем памяти анализатора не превышал 400 Кбайт.

- повышение покрывающей способности словарей происходит путем включения в состав словарного комплекса словарей словоизменительных парадигм слов, представлять которые будут их нормальные формы;
- сокращение трудозатрат на разработку программных средств и повышение их быстродействия решается за счет разработки относительно простых алгоритмов с небольшой вычислительной сложностью.

Учитывая, что реализация функциональных требований в значительной степени зависит от состава и структуры используемых словарей, дополнительно нами были разработаны частные требования к их составу и структуре, которые обеспечивают:

1) Словарь *S* – назначение полного набора грамматических и семантических характеристик наиболее частотным словоформам;

2) Словарь *K* – назначение усеченного набора грамматических характеристик словоформам с аномальной трансформационной системой словоизменения;

3) Словарь *E* – назначение усеченного набора грамматических характеристик словоформам с регулярной трансформационной системой словоизменения;

4) Словарь *C* – назначение семантических признаков всем членам словоизменительных парадигм слов;

5) Таблица *N* – преобразование текстовой словоформы в ее нормальную форму;

6) Таблица *T* – преобразование усеченного набора грамматических характеристик в их полный набор.

Рассмотрим назначение, лексический состав и набор грамматических и семантических характеристик указанных типов словарей и грамматических таблиц.

Словарь *S* предназначен для обработки основного высокочастотного потока текстовых словоформ. Он включает как все служебные слова, так и наиболее часто встречающиеся формы слов. Обеспечивает назначение набора грамматических и семантических признаков, представляющего более 30 возможных характеристик. Формат словаря – FMA_s30. В табл. 1 приводится фрагмент словаря *S*.

Словарь *K* разработан для обработки остального потока текстовых словоформ. В него включены формы слов, относящиеся к аномальной трансформационной системе словоизменения и словообразования, а также часто встречающиеся формы слов русского языка. Обеспечивает назначение усеченного набора грамматических признаков, состоящего из пяти возможных характеристик. Формат словаря – FMA_k05. В табл. 2 приводится фрагмент словаря *K*.

Словарь *E* используется для обработки словоформ, относящихся к регулярной трансформационной системе словоизменения и словообразования слов русского языка, не покрытых лексикой словарей *S* и *K*. В его состав входят конечные буквосочетания форм слов. Делает возможным назначение усеченного набора грамматических признаков, состоящего из пяти возможных характеристик. .Формат словаря – FMA_k05. В табл. 3 приводится фрагмент словаря *E*.

Словарь *C* создан для обработки представителей словоизменительных парадигм слов и предоставляет возможность назначения всем их членам набора семантических признаков, а также обеспечивает назначение набора словообразовательных и семантических характеристик, состоящего из 18 возможных характеристик. Формат словаря – FMA_c18. В табл. 4 приводится фрагмент словаря *C*.

Таблица *N* служит для преобразования текстовой формы слова в его нормальную форму и содержит нормализующие окончания слов. Включает следующие грамматические признаки: номер флективного класса словоформы и нормализующее окончание, соответствующее номеру флективного класса. Формат таблицы – FMA_n02. В табл. 5 продемонстрирован её фрагмент.

Таблица *T* предназначена для преобразования усеченного набора грамматических признаков словоформ в его полный состав. Обеспечивает назначение полного набора формообразующих характеристик, включающего более восьми возможных характеристик. Формат таблицы – FMA_t08. В табл. 6 приводится её фрагмент.

Таблица 1

Фрагмент словаря *S*

автор	#OK=0#FK=21#GI=*1110#GK=N #OS=QA#TW=1#LI =1#TK=k#RW =w#PD=t#TD=S
администрации	#OK=1#FK=61#GI=*2120*2130*2160*2210*2240#GK=N#SU=t#OS=xf#TK=k#TW=1#TD=S
вашего	#OK=3#FK=114#GI=*1120*1140*3120#GK=m#LI=2#RW=w#OS=QA#TW=2#LI=3#TK=k #PD=t#TK=v #TD=S
кремлевские	#OK=2#FK=106#GI=*0210*0240#GK=A#SU=t#OS=Цg#TW=1#LI=2#TK=k #RW=w#TD=S

Таблица 2

Фрагмент словаря *K*

автоматы	#OK=1#FK=1#TW=1#RW=w#TD=K
автомашину	#OK=1#FK=56#TW=1#RW=w#TD=K
автомеризация	#OK=1#FK=61#TW=1#RW=w #TD=K
автомеханик	#OK=0#FK=31#TW=1#RW=w #TD=K

Таблица 3

Фрагмент словаря *E*

аболз	#OK=1#FK=56#TW=1#RW=w#TD=E
абонзо	#OK=1#FK=1#TW=1#RW=w#TD=E
абор	#OK=1#FK=56#TW=1#RW=w#TD=E
аборок	#OK=1#FK=1#TW=1#RW=w#TD=E

Таблица 4

Фрагмент словаря *C*

автомеханик	#PD=t #TD=C
весь	#DK=e#TD=C
который	#DK=k#TD=C
одна	#DK=0#TD=C
может	#DK=M#TD=C
август	#PT=t#TD=C

Таблица 5

Фрагмент таблицы *N*

#FK=107	#NO=ой
#FK=110	#NO=ой
#FK=111	#NO=ий
#FK=112	#NO=+
#FK=113	#NO=й

Таблица 6

Фрагмент таблицы *T*

#TO=+#FK=011	#GI=*1110*1140#OS=IA#SU=t#GK=N#TD=T
#TO=+#FK=014	#GI=*1110*1140#OS=LA#SU=t#GK=N#TD=T
#TO=+#FK=015	#GI=*1110*1140*1220#OS=MA#SU=t#GK=N#TD=T
#TO=+#FK=016	#GI=*1110*1140#OS=NA#SU=t#GK=N#TD=T
#TO=+#FK=017	#GI=*1110*1140*1220#OS=OA#SU=t#GK=N#TD=T

АЛГОРИТМ МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА МЕТАФРАЗ ВТОРОГО ПОКОЛЕНИЯ

Обработка словоформ алгоритмом морфологического анализатора МетаФраз второго поколения выполняется следующим образом: вначале производится прямой поиск анализируемой словоформы по ее буквенному коду в словаре *S* и, если она там находится, ей назначается полный набор грамматических и семантических характеристик и анализ этой словоформы заканчивается. Если эта словоформа не была обнаружена в словаре *S*, то производится прямой поиск по ее буквенному коду в словаре *K*. В случае обнаружения в этом словаре словоформе назначается усеченный набор грамматических характеристик и далее осуществляется преобразование набора усеченной грамматической информации в полный ее состав по таблице *T*. Потом производится назначение анализируемой словоформе семантических признаков по словарю *C*, но предварительно эта словоформа должна быть приведена к ее нормальной форме по таблице *N*.

Словоформы, не обнаруженные в словарях *S* и *K*, обрабатываются по методу лингвистической аналогии на основе анализа их конечных буквосочетаний. Для этого выполняются инверсия буквенного состава словоформы и поиск на наибольшее совпадение ее конечного буквосочетания с буквосочетанием одного из элементов словаря *E*. После нахождения такого буквосочетания анализируемому слову назначается усеченный набор грамматических характеристик. Дальнейшая обработка производится по схеме, аналогичной обработке словоформ, найденных в словаре *K*.

Приведем алгоритм МетаФраз второго поколения.

Алгоритм морфологического анализатора МетаФраз второго поколения

Шаг 1. Выполняется поиск анализируемой формы слова на полное ее совпадение в словаре *S*. В случае успешного поиска словоформе назначается грамматическая и семантическая информация (в соответствии с форматом FMA_s30) и выполняется переход к шагу 7. В случае отсутствия этой словоформы в словаре – переход к шагу 2.

Таблица 7

Результаты работы морфоанализаторана основе технологий МетаФраз второго поколения

00 По	#OK=0#FK=156#GI=*0030#GK=F#OS=ыА#TW=1#TD=S
01 данным	#OK=2#FK=103#GI=*0230*1150*3150#GK=A#OS=ФУ#TW=1#TD=S
02 флота	#OK=1#FK=1#GI=*1120#GK=N#OS=АВ#TW=1#TD=S
03 ,	#OK=0#FK=0#TW=1#GI=*0000#OS=00#GK=,
04 в	#OK=0#FK=164#GI=*0040*0060#GK=F#OS=1А#TW=1#TD=S
05 уходящем	#OK=2#FK=105#TW=1#TD=E#GI=*1160*3160#OS=8S#GK=A#TD=T
06 году	#OK=1#FK=10#GI=*1130*1160#GK=N#OS=Н3#TW=1#TD=S
07 было	#OK=1#FK=125#GI=*3100#GK=L#PG=t#OS=еу#TW=1#TD=S
08 проведено	#OK=1#FK=126#TW=1#TD=K#GI=*3100#OS=еу#GK=K#TD=T#PG=t#TD=C
09 более	#OK=0#FK=152#TW=1#TD=K#GI=*0000#OS=шА#GK=Y#TD=T
10 150	#OK=0#FK=145#GI=*1000#GK=0#OS=уА#TW=1#TD=S
11 учений	#OK=1#FK=73#GI=*3220#GK=N#OS=Йх#TW=1#TD=S
12 различной	#OK=2#FK=103#TW=1#TD=E#GI=*2120*2130*2150*2160#OS=ФВ#GK=A#TD=T
13направленности	#OK=1#FK=55#TW=1#TD=E#GI=*2120*2130*2160*2210*2240#OS=tf#SU=t#GK=N#TD=T
14 ,	#OK=0#FK=0#TW=1#GI=*0000#OS=00#GK=,
15 в	#OK=0#FK=164#GI=*0040*0060#GK=F#OS=1А#TW=1#TD=S
16 ходе	#OK=1#FK=1#GI=*1160#GK=N#OS=АК#TW=1#TD=S
17 которых	#OK=2#FK=103#GI=*0220*0240*0260#GK=k#OS=ФХ#TW=1#TD=S
18 выполнено	#OK=1#FK=126#TW=1#TD=E#GI=*3100#OS=еу#GK=K#TD=T#PG=t#TD=C
19 свыше	#OK=0#FK=155#TW=1#TD=K#GI=*0020#OS=ъА#GK=F#TD=T
20 500	#OK=0#FK=145#GI=*1000#GK=0#OS=уА#TW=1#TD=S
21 боевых	#OK=2#FK=107#GI=*0220*0240*0260#GK=A#OS=ЧХ#TW=1#TD=S
22 упражнений	#OK=1#FK=73#TW=1#TD=E#GI=*3220#OS=Йх#GK=N#TD=T
23 и	#OK=0#FK=153#GI=*0000#GK=&#OS=ЩА#TW=1#TD=S
24 применений	#OK=1#FK=73#TW=1#TD=E#GI=*3220#OS=Йх#GK=N#TD=T
25 оружия	#OK=1#FK=73#GI=*3120*3210*3240#GK=N#SU=t#OS=Йм#TW=1#TD=S
26 .	#OK=0#FK=0#TW=1#GI=*0000#OS=00#GK=.

Шаг 2. Выполняется поиск анализируемой словоформы на полное ее совпадение в словаре *K*. В случае успешного поиска ей назначается усеченная грамматическая информация (в соответствии с форматом FMA_k5) и выполняется переход к шагу 4. В случае отсутствия этой словоформы в словаре – переход к шагу 3.

Шаг 3. Производится инверсия буквенного состава анализируемого слова и выполняется поиск конечного буквосочетания анализируемой формы слова на наибольшее совпадение с одним из элементов словаря *E*. В случае успешного поиска словоформе назначается усеченная грамматическая информация (в соответствии с форматом FMA_k5) и выполняется переход к шагу 4.

Шаг 4. Выполняется поиск в таблице *T* по двум грамматическим характеристикам – номеру флективного класса и текстового грамматического окончания. В случае успешного поиска словоформе назначается полный набор грамматической информации (в соответствии с форматом FMA_t8) и выполняется переход к шагу 5.

Шаг 5. По таблице *N* выполняется приведение текстовой словоформы к ее нормальной форме путем присоединения к словоизменительной основе нормализующего окончания, соответствующего номеру флективного класса(в соответствии с форматом FMA_n2). По завершению операции выполняется переход к шагу 6.

Шаг 6. Выполняется прямой поиск сформированной на шаге 5 нормальной формы слова на полное его совпадение в словаре *C*. В случае успешного поиска всем членам словоизменительной парадигмы назначается семантическая информация (в соответствии с форматом FMA_c18) и выполняется переход к шагу 7. В случае отсутствия представителя словоизменительной парадигмы в словаре – переход к шагу 7.

Шаг 7. Выполняется преобразование полученных результатов в структуру метаданных.

В табл. 7 приводятся результаты работы анализатора на основе технологий МетаФраз второго поколения.

ХАРАКТЕРИСТИКИ ПРОГРАММНЫХ И ДЕКЛАРАТИВНЫХ СРЕДСТВ

При проектировании анализатора на основе технологий МетаФраз второго поколения должны быть предварительно установлены параметры вычислительной сложности алгоритмов[3] отдельных процедур с целью определения их быстродействия при реализации программного кода. Для упрощения этой зада-

[3]Под термином «вычислительная сложность алгоритма» в информатике и теории алгоритмов понимается функция зависимости объема работы, которая выполняется некоторым алгоритмом, от размера входных данных. Объем работы обычно измеряется абстрактными понятиями времени и пространства, называемыми вычислительными ресурсами.

чи за единицу вычислительной сложности процедуры примем прямой поиск в хешированном массиве, при этом сопутствующие вычислительные действия, связанные с преобразованием данных, также включим в эту единицу сложности вычислений. Операцию обратного поиска на наибольшее вхождение оценим в пять единиц.

Суммарные результаты вычислительной сложности каждой технологической цепочки операций морфоанализатора МетаФраз первого поколения приведены в табл. 8, а в табл. 9 – аналогичные результаты вычислительной сложности каждой технологической цепочки операций морфоанализатора на основе технологий МетаФраз второго поколения.

Анализ табл. 8 и 9 показывает, что минимальная вычислительная сложность технологической цепочки №1 МетаФраз первого поколения равна двум, а цепочки №2 – семи. Для анализатора на основе технологий МетаФраз второго поколения этот диапазон значений более широкий. Так, вычислительная сложность технологической цепочки №1 анализатора здесь равна единице, вычислительная сложность технологической цепочки №2 – пяти, а вычислительная сложность технологической цепочки №3 – 10.

Таблица 8

Суммарные результаты вычислительной сложности каждой технологической цепочки операций в морфоанализаторе МетаФраз первого поколения*

Технологич. цепочки операций	Поиск в словарях и грамматических таблицах			Суммарная вычислительная сложность технологических цепочек операций
	КСС	КБС	ФКГИ	
№1	1	–	1	2
№2	1	5	1	7

* 1. Технологической операцией цепочки №1 является а) прямой поиск в словаре коротких и служебных слов (КСС) и б) поиск в словаре флективный класс – грамматическая информация (ФКГИ).

2. Технологической операцией цепочки №2 является а) прямой поиск в словаре коротких и служебных слов (КСС), б) обратный поиск в словаре конечных буквосочетаний и в) поиск в словаре флективный класс – грамматическая информация (ФКГИ).

Таблица 9

Суммарные результаты вычислительной сложности каждой технологической цепочки операции в морфоанализаторе МетаФраз второго поколения*

Технологич. цепочки операций	Поиск в словарях и грамматических таблицах						Суммарная вычислительная сложность технологических цепочек операций
	S	*K*	*E*	*T*	*N*	*C*	
№1	1	–	–	–	–	–	1
№2	1	1	–	1	1	1	5
№3	1	1	5	1	1	1	10

***** 1. Технологической операцией цепочки №1 является а) прямой поиск в словаре S.

2. Технологической операцией цепочки №2 является а) прямой поиск в словаре S, б) прямой поиск в словаре K, в) прямой поиск в таблице T, д) прямой поиск в таблице N, г) прямой поиск в словаре C

3. Технологической операцией цепочки №3 является а) прямой поиск в словаре S, б) прямой поиск в словаре K, в) обратный поиск в словаре E, д) прямой поиск в таблице T, г) прямой поиск в таблице N, е) прямой поиск в словаре C.

Таблица 10

Параметры комплекса словарей МетаФраз первого поколения

Тип словаря	Объем словаря	Эффективный объем словоформ	Количество граммат. признаков	Вероятность правильного назначения грамматической информации	Генерация словоформ в текстах сверхбольшого объема	Покрытие текстов каждым словарем, %
Словарь коротких и служебных слов	78000	78000	4	100%	13731372	53
Словарь конечных буквосочетаний	44000	3–7 млн	4	77%	1–3 млн	100

Таблица 11

Параметры комплекса словарей МетаФраз второго поколения

Тип словаря	Объем словаря	Эффективный объем словоформ	Количество грамматич. и семантич. признаков	Вероятность правильного назначения грамматической информации, %	Встречаемость словоформ в текстах большого объема	Покрытие текстов каждым словарем, %
Словарь S	40000	40000	34	100	26078948	81
СловарьK	120000	120000	8	100	27431372	97
Словарь E	150000	3–20 млн	8	87	3–20 млн	100
Словарь C	120000	600000	26	100	27414081	99

Очевидно, что вычислительная сложность алгоритмов и процедур МетаФразв значительной степени зависит от соотношения объемов текстовых слов, которые будут обрабатываться каждой технологической цепочкой операций. Поэтому увеличение потока текстовых слов, обрабатываемых технологической цепочкой №1, обеспечит снижение общей сложности обработки процедур МетаФраз и, соответственно, приведет к увеличению ее быстродействия. Таким образом, видно, что количественные параметры комплекса словарей МетаФраз напрямую влияют на их быстродействие.

Ниже приведены количественные параметры комплекса словарей МетаФраз первого (табл. 10) и второго (табл. 11) поколений, иллюстрирующие динамику изменения и перераспределения количественного состава словарей МетаФраз.

ТЕХНОЛОГИИ СОЗДАНИЯ И ВЕДЕНИЯ ДЕКЛАРАТИВНЫХ СРЕДСТВ

Очевидно, что создание необходимых объемов разного типа словарей и большое количество сопутствующих грамматических и семантических характеристик их элементов невозможно реализовать ручными методами. Для этого требуются программно-лингвистические средства автоматизации создания словарей и грамматических таблиц, которые в рамках концепции фразеологического анализа текстов принято называть Автоматизированной словарной службой[4] (АСС). Кратко определим объекты и средства автоматизации в рамках АСС.

Под *объектами автоматизации* АСС будем понимать массивы групп и подгрупп словоформ русского языка, расклассифицированные по совокупности грамматических и семантических характеристик. Общее число таких групп и подгрупп словоформ будет составлять несколько десятков.

Средствами автоматизации АСС обозначим технологические операции, приводящие к трансформации их буквенного кода и состава грамматических и семантических признаков.

Каждое текущее значение состава грамматических и семантических признаков назовем *форматом словоформы*.

К *технологической операции* АСС отнесем локальное изменение формата словоформы из одного состояния в другое.

Таким образом, основная задача реализации *технологий АСС* – это автоматизированное выполнение цепочки технологических операций с целью *создания комплекса словарей для морфологического анализа МетаФраз второго поколения*, при минимальном участие человека в этом процессе.

Источниками информации для формирования декларативных средств могут служить следующие лингвистические ресурсы.

1. Для формирования словаря *S* источником являются частотные словари словоформ, созданные на больших корпусах текстов. Необходимо обработать и включить в состав словаря *S* частотную часть этих словарей, статистическая информация об одном из которых содержится в табл. 12.

2. Для формирования словаря *K* источником служит словарь коротких и служебных слов МетаФраз первого поколения. Необходимо будет выполнить переформатирование грамматических характеристик словаря.

3. Источник информации для формирования словаря *E* – словарь конечных буквосочетаний МетаФраз первого поколения. Необходимо будет выполнить переформатирование грамматических характеристик словаря.

4. Словарь *C* формируется на основе частотных словарей представителей словоизменительных парадигм, созданных на больших корпусах текстов. Необходимо обработать и включить в состав словаря *C* частотную часть этих словарей, статистическая информация об одном из которых содержится в табл. 13.

Основные технологические операции Автоматизированной словарной службы

1. Назначение базовых грамматических характеристик словоформам на основе методов лингвистической аналогии.

2. Автоматическое назначение полного набора грамматических характеристик на основе анализа их базовых характеристик.

[4]Автоматизированная словарная служба (АСС) – это сложный программно-информационный комплекс, обеспечивающий возможность автоматизированной интеллектуальной обработки текстовой информации с целью ее преобразования в систему словарных конструкций, сопровождаемых совокупностью их грамматических и семантических характеристик.

3. Вычисление совокупности одних семантических характеристик на основе анализа других характеристик.

4. Шаблонное назначение семантических характеристик, присущих конкретной группе слов.

Как отмечалось, основным источником пополнения словарей *S* и *C* была лексика тематических корпусов текстов, которая выбиралась по статистическим параметрам. Для получения этих параметров были сформированы частотные словари двух типов: словарь словоформ и словарь словоизменительных парадигм. В частотном словаре словоизменительных парадигм каждая парадигма представлена нормализованной формой слова парадигмы.

Таблица 12

Статистические данные о частотном словаре словоформ, составленном по корпусу текстов общим объемом 28,5 млн слов

Ранги частот	Макс.частота диапазона словоформ	Мин.частота диапазона словоформ	Количество разных словоформ в корпусе текстов	Общее количество словоформ в корпусе текстов	Покрытие корпуса текстов словоформами
1	1163633	22988	100	9033294	0,316575
2	22988	5696	500	12977269	0,454793
3	5696	3140	1000	15064240	0,527932
4	3140	1645	2000	17286279	0,605804
5	1645	1110	3000	18625593	0,652740
6	1110	833	4000	19583726	0,686319
7	833	633	5000	20329325	0,712448
8	633	303	10000	22547206	0,790175
9	303	185	15000	23731372	0,831674
10	185	127	20000	26422951	0,858814
11	127	73	30000	25474618	0,892767
12	73	48	40000	26078948	0,913946
13	48	11	50000	27161190	0,927571
14	11	11	100000	27500566	0,963767
15	11	4	194461	28046791	0,982910
16	4	2	305591	27552620	0,992041
17	2	1	532706	28534454	1,000000

Количество разных словоформ в словаре равно 532 706.

Таблица 13

Статистические данные о частотном словаре нормализованных форм слов словоизменительных парадигм, составленном по корпусу текстов общим объемом 28,5 млн слов

Ранги частот	Макс. частота диапазона словоизм. парадигм	Мин. частота диапазона словоизм. парадигм	Количество разных словоизм. парадигм в корпусе текстов	Обще количество словоизм. парадигм в корпусе текстов	Покрытие корпуса текстов словоизм. парадигм
1	1164325	30960	100	10560385	0,383281
2	30960	7113	500	15848771	0,575218
3	7113	3743	1000	18385769	0,667297
4	3743	1752	2000	20900122	0,758553
5	1752	1068	3000	22254899	0,807724
6	1068	726	4000	23132119	0,839562
7	726	532	5000	23753139	0,862101
8	532	191	10000	25340828	0,919725
9	191	99	15000	26032609	0,944832
10	99	60	20000	26422951	0,959000
11	60	28	30000	26840747	0,974163
12	28	16	40000	27053112	0,981871
13	16	11	50000	27161190	0,985793
14	11	3	100000	27414081	0,994972
15	3	2	125444	27468955	0,996963
16	2	1	209109	27552620	1,000000

Количество разных словоизменительных парадигм слов в словаре равно 209 109.

Словарь первого типа служил источником для пополнения словаря *S*. Словарь второго типа – источником для пополнения словаря *C*. Статистические данные о частотном словаре форм слов корпуса текстов, общим объемом 28 534 454 слов, приведены в табл. 12. Статистические данные о частотном словаре представителей словоизменительных парадигм этого корпуса текстов приведены в табл. 13.

На основе имеющихся статистических данных о конкретных словоформах (табл. 12) и статистических данных о нормализованных формах слов, представляющих словоизменительные парадигмы (табл. 13), возможно принимать решения по формированию и пополнению словарей *S* и *C* частотной лексикой.

ЗАКЛЮЧЕНИЕ

По итогам создания морфологического анализатора на основе технологий МетаФраз второго поколения можно отметить следующее.

1. В основу анализатора была положена морфологическая модель, основанная на системе флективных классов русского языка.

2. Система флективных классов, разработанная профессором Г.Г. Белоноговым в конце 60-х гг. прошлого века, представляет многообразие типов словоизменения русского языка.

3. Принцип лингвистической аналогии дает возможность обеспечить назначение грамматических характеристик словоформ бессловарным методом, основанным на анализе конечных буквосочетаний слов, а также позволяет сократить трудозатраты при формировании декларативных средств.

4. Повышение быстродействия предлагаемого морфоанализатора за счет включения в его состав дополнительной быстродействующей процедуры, обрабатывающей основной поток (не менее 80%) текстовых словоформ.

5. Повышение качества декларативных средств путем расширения спектра грамматических (формообразующих и словообразующих) и семантических характеристик.

6. Автоматизация назначения грамматических и семантических характеристик как словоформам, так и нормальным формам слов.

7. Увеличение покрывающей способности словарей при включении в состав словарного комплекса наряду со словарями словоформ словарей словоизменительных парадигм слов в виде их нормальных форм.

8. Сокращение трудозатрат на разработку программных средств и повышение их быстродействия путем разработки относительно простых алгоритмов с небольшой вычислительной сложностью.

В заключение определим назначение и местоположение в составе базовых средств семантико-синтаксического анализа текстов процедуры разрешения омонимии[5]. На наш взгляд разрешение как лексической (частеречевой), так и семантической омонимии невозможно выполнить без учета контекста. Между тем процедура морфологического анализа ориентирована на анализ слов вне контекста. Поэтому в предлагаемом морфоанализаторе МетаФраз второго поколения эта операция не предусмотрена. Ее реализация возможна только на этапе семантико-синтаксического или концептуального анализа [2, 6], когда появляется возможность опираться на контекст омонимичной формы слова, но при этом предварительно морфологический анализатор должен предоставить информацию о наличии омонимии у анализируемой формы слова.

СПИСОК ЛИТЕРАТУРЫ

1. Белоногов Г.Г., Гиляревский Р.С., Селедков С.Н., Хорошилов Ал-р А. О путях повышения качества поиска текстовой информации в системе Интернет // Научно-техническая информация. Сер. 2. – 2013. – № 8. – С. 15–22; Belonogov G.G., Gilyaresvkii R.S., Seletkov S.N., Khoroshilov A.A. Ways to Improve the Quality of Textual Data Searches on the Internet. – Automatic Documentation and Mathematical Linguistics. – 2013. – Vol. 47, № 4. – P. 111-120.
2. Аблов И.В., Козичев В.Н., Ширманов А.В., Хорошилов Ал-р А., Хорошилов Ал-ей А. Средства машинной грамматики русского языка (по Г.Г. Белоногову) // Научно-техническая информация. Сер. 2. – 2018. – № 6. – С. 32-46.
3. Белоногов Г.Г., Калинин Ю.П., Хорошилов Ал-др А., Хорошилов Ал-ей А. Компьютерная лингвистика и перспективные информационные технологии // Научно-техническая информация. Сер. 2. – 2004. – № 8. – С. 22–32.
4. Старовойтов А.В., Пошатаев О.Н., Прохоров С.Н., Хорошилов Ал-р А. Методы автоматизированного составления и ведения словарей // Информатизация и связь. – 2013. – №3. – С. 91–97.
5. Белоногов Г.Г., Зеленков Ю.Г., Новоселов А.П., Хорошилов Ал-др А., Хорошилов Ал-ей А. Метод аналогии в компьютерной лингвистике // Научно-техническая информация. Сер. 2. – 2000. – № 1. – С. 21–31.
6. Кан А.В., Ревина В.Д., Руснак В.И., Хорошилов Ал-др А., Хорошилов Ал-сей А. Автоматическоеформирование синтаксической модели языка для задач машинного перевода и информационного поиска // Научно-техническая информация. Сер. 2. – 2018. – № 12. – С. 25-41.

Материал поступил в редакцию 08.02.21.

[5] Академик В.В. Виноградов в статье «Об омонимии и смежных с ней явлениях» (журнал «Вопросы языкознания» 1968 г.) определил это лингвистическое понятие как «..звуковое и грамматическое совпадение языковых единиц, которые семантически не связаны друг с другом..»

Сведения об авторах

ХОРОШИЛОВ Александр Алексеевич – доктор технических наук, профессор НИУ МАИ; ведущий научный сотрудник Федерального исследовательского центра Информатики и Управления (ФИЦ ИУ) РАН; старший научный сотрудник 27 ЦНИИ Министерства обороны РФ, Москва
e-mail: khoroshilov@mail.ru

НИКИТИН Юрий Викторович – научный сотрудник ФИЦ ИУ РАН, руководитель группы разработки АО «НПК «ВТ и СС»
e-mail: yuri.v.nikitin@gmail.com

ПШЕНИЧНЫЙ Сергей Игоревич – кандидат экономических наук, директор программ АО «НПК «ВТ и СС»
e-mail: s.pshenichniy@htsts.ru

ШЕВКУНОВ Максим Александрович – ведущий конструктор АО «НПК «ВТ и СС»
e-mail: mshevkunov@htsts.ru

ХОРОШИЛОВ Алексей Алексеевич – кандидат технических наук, старший научный сотрудник 27 ЦНИИ Министерства обороны РФ, Москва
e-mail: alex_khoroshilov@mail.ru

УДК 811.351.42’322.6

Э.С. Израилова, А.С. Бадаева

Анализ качества речевого сигнала системы синтеза чеченской речи

Описывается создание прототипа системы синтеза чеченской речи, основанного на нейросетевой модели DCTTS и состоящего из различных функциональных модулей. Анализируется качество полученного речевого сигнала в соответствии с международной оценкой качества синтезируемой речи MOS, а также результаты тестирования прототипа системы синтеза чеченской речи и его доработки. Представлены новые направления исследований, предварительно ориентированные на устранение графической омонимии в тексте.

Ключевые слова: *система синтеза чеченской речи, обучающий корпус, анализ качества синтезируемой речи, MOS-оценка, графическая омонимия*

DOI: 10.36535/0548-0027-2021-04-4

Разработка системы синтеза чеченской речи началась с изучения актуальных методов и систем синтеза речи наиболее популярных языков мира, по которым ученым удалось достигнуть высоких результатов в области речевых технологий. Основное внимание в настоящей работе было уделено популярным в последние годы нейросетевым моделям и алгоритмам синтеза и распознавания речи.

РАЗРАБОТКА ПРОТОТИПА СИСТЕМЫ СИНТЕЗА ЧЕЧЕНСКОЙ РЕЧИ

Практическая работа по проекту синтеза чеченской речи была связана с подготовкой обучающего корпуса, основу которого составили тексты на чеченском языке, структурированные в виде отдельных пронумерованных предложений и соответствующих им фонограмм речи. Созданная таким образом база данных предназначена для машинного обучения нейросетевых систем синтеза и распознавания речи и была использована нами при разработке системы синтеза чеченской речи. Этот прототип системы синтеза чеченской речи состоит из различных функциональных модулей: модуль транскрибирования чеченских текстов, нормализатор для расшифровки числительных и сокращений; модуль обучения, включающий две нейронные сети; модуль синтезирования речи (на основе вокодера).

Создание транскрипций осуществлено при помощи программы автоматического транскрибирования чеченских текстов Elp-Az, в которой используются латинские буквы и символы из фонетического алфавита AZBAT, ранее разработанного нами в качестве основы соответствующего модуля будущей системы синтеза чеченской речи [1].

Для моделирования системы синтеза чеченской речи было решено остановиться на архитектуре глубоких сверточных нейронных сетей DeepConvolutionalTextToSpeech (DCTTS), как наиболее оптимальной по соотношению время обучения/качество синтеза. Модель DCTTS демонстрирует высокую производительность и скорость обучения и имеет относительно ограниченные требования к вычислительной мощности компьютера [2, 3].

Обучение прототипа системы продолжалось 26 часов, было выполнено 500 тыс. итераций для нейронной сети Text2Mel, 240 тыс. итераций для сети SSRN [4]. Процесс обучения обеспечивал компьютер с двумя графическими процессорами, видеопамятью емкостью 12 Гб и оперативной памятью – 64 Гб. В результате нам удалось синтезировать чеченскую речь.

АНАЛИЗ КАЧЕСТВА ПОЛУЧЕННОГО РЕЧЕВОГО СИГНАЛА

Для объективной оценки качества синтезированной речи существует несколько методик. В России чаще всего используется ГОСТ Р 50840-95 «Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости», а также различные тесты отдельных компонентов, но единого стандарта оценки пока нет. В мировой практике разработчики речевых приложений ориентируются на рекомендации P.85 ITU-T «Метод субъективной оценки качества речи устройств речевого вывода».

В сфере речевых технологий проводятся различные конкурсы, например, международный конкурс Blizzard Challenge – «соревнования» синтезаторов. Голоса для сравнения систем синтеза создаются на основе одних и тех же звуковых баз данных, предоставляемых перед началом соревнований. По истечении времени, отведенного на создание голосов, участникам конкурса выдается набор текстов, синтезированные звуковые файлы для которых необходимо предоставить организаторам для оценки. В 2010 г. соревнования проводились для корпусов речи на английском и

китайском языках. В 2012 г. в качестве дополнительного задания предлагалось разработать собственный метод оценки качества синтеза и провести оценку. По образцу Blizzard Challenge для испаноязычных синтезаторов был организован конкурс Albayzin. Существует также стандартизованный набор тестов для синтеза речи на французском языке, разработанный в ходе национального проекта EvaSy (Evaluation of speech synthesis systems – оценка систем синтеза речи) [5].

Один и тот же текст можно прочитать бесконечным количеством способов – единого способа для произношения конкретной фразы не существует. Поэтому часто оценки качества синтеза речи субъективны и зависят от восприятия слушающего. Методы оценки можно разделить на две большие группы: субъективные (MOS-оценка) и инструментальные.

Стандартные критерии анализа качества синтезированной речи – это MOS (Mean Opinion Score), усредненная оценка качества и естественности речи, выданная слушателями для синтезированных аудио по шкале от 1 до 5. Единица означает совсем неправдоподобное звучание, а пятерка – речь, неотличимую от человеческой. Реальные записи голоса обычно получают значения примерно 4,5, и значение больше 4-х считается достаточно высоким.

В рамках данного проекта для анализа качества синтезируемой чеченской речи была использована MOS – оценка по пятибалльной шкале по нескольким категориям: общее впечатление, слуховое усилие, естественность, понимание смысла сообщения, темп, разборчивость, приятность голоса. Качество речи оценивалось носителями синтезируемого языка. На сайте отдела прикладной семиотики Академии наук Чеченской республики (ps95.ru) было проведено онлайн тестирование синтезированных системой предложений (рис. 1). Тест представлял собой опросник, заполняемый интернет-пользователями. При разработке опросников использовались рекомендации P.85 ITU-T «Метод субъективной оценки качества речи устройств речевого вывода».

Для оценки качества было выбрано 10 стилистически различных предложений (проза, диалог, отрывок стихотворения). Прототипом системы синтеза чеченской речи был выполнен синтез этих предложений. Озвученные образцы речи были включены в тест-опросник и выставлены на сайте. После прослушивания каждого предложения респонденту предоставлялась возможность выбора расслышанных им слов в предложении. Хотя такой вид тестирования не рекомендован ГОСТом и системой оценки качества MOS, мы включили и такой способ оценки, так как, по нашему мнению, этот метод дает дополнительную информацию относительно разборчивости речи.

Тест был подготовлен с помощью online-сервиса «Google Формы» и выставлен на сайте www.dosh-speech.ru. По социальным сетям была отправлена ссылка на опрос и обращение принять участие в тестировании. В опросе приняли участие 74 респондента. Затем нами были проанализированы ответы и выявлены следующие показатели качества синтезируемых аудио-образцов.

1. В горизонтальных гистограммах на рис. 2 отображена разборчивость речи каждого из десяти предложений – она составляет 85,2%.

2. Качество синтезируемой речи приближено к естественной MOS-оценке речи (от 4 до 5). Оценка различных характеристик речи, отображена в гистограммах на рис. 3, в которых большая часть ответов респондентов составляют оценки 4 (рис. 3).

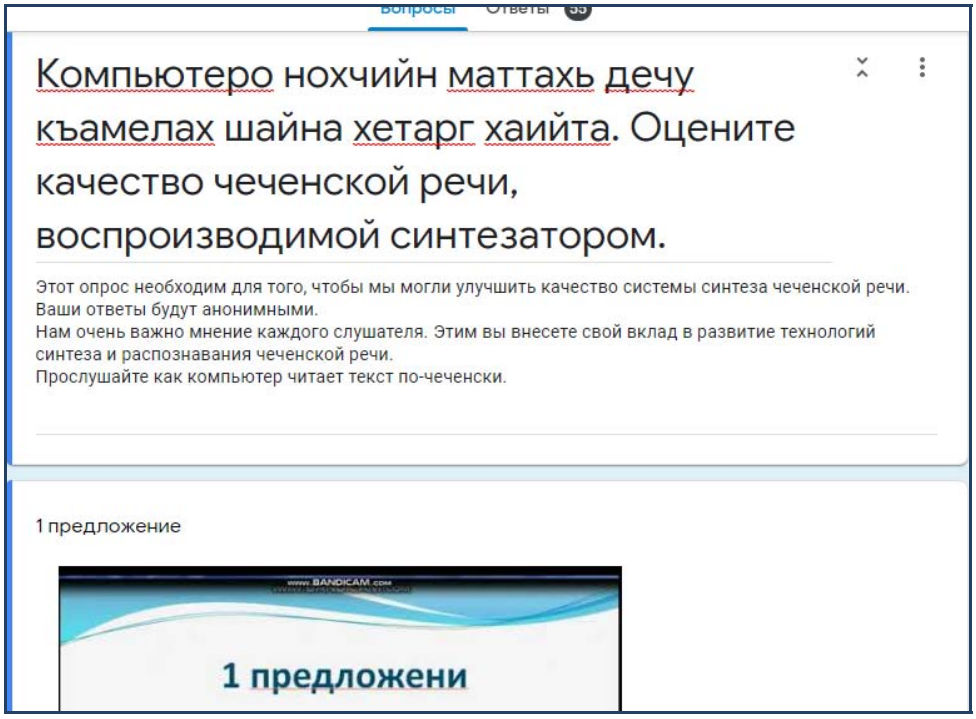


Рис. 1. Тестирование синтезированных предложений

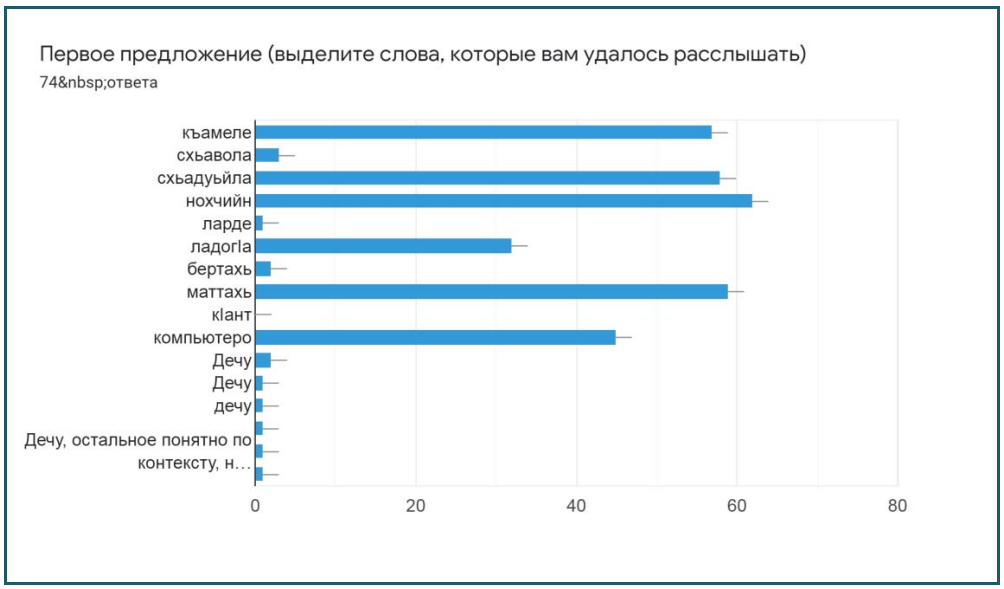


Рис. 2. Оценка разборчивости синтезируемой речи

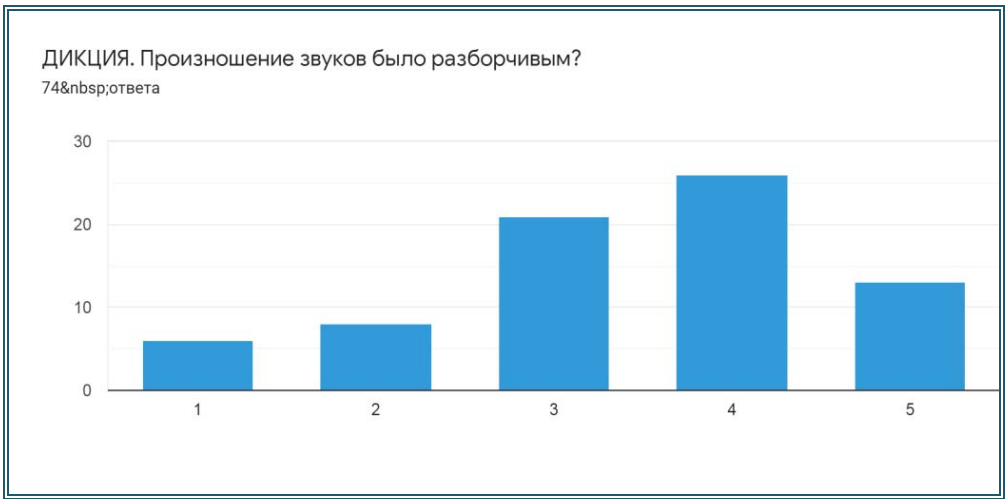


Рис. 3. Оценка качества синтезируемой речи

АНАЛИЗ ТЕСТИРОВАНИЯ ПРОТОТИПА СИСТЕМЫ СИНТЕЗА ЧЕЧЕНСКОЙ РЕЧИ

После анализа качества и разборчивости синтезируемой прототипом системы синтеза чеченской речи были выявлены как достоинства созданного нами приложения, так и его недостатки.

Минусы прототипа системы синтеза чеченской речи.

1. В качестве одной из основных проблем при реализации проекта с самого начала рассматривалось неумение синтезатора разграничивать долготу / краткость гласных, а также дифтонг / монофтонг (иэ / э, уо / о), поскольку в чеченской графике отсутствуют специальные маркеры этих признаков на письме. Особенно трудно решаема эта проблема для омографов, имеющих различное произношение одних и тех же гласных. Если в других случаях эту проблему можно частично снять, по некоторым формальным признакам определив категориальную принадлежность словоформ, то с омографами, естественно, такой подход не работает.

2. Несколько неожиданной для разработчиков оказалась еще одна проблема – это чтение слов, заимствованных из русского языка или через русский язык. С другими заимствованиями, более ранними, подобные проблемы отсутствуют, поскольку они уже «адаптировались» к чеченскому языку и своим звучанием не отличаются от чеченских слов. Несколько иначе обстоит дело с заимствованиями из русского языка, которые вошли в чеченскую речь в XX в. Эти слова заимствованы практически без изменений звукового состава и поэтому они, как правило, значительно выделяются на общем фоне чеченской речи. Однако компьютерная система, обученная по принципу частотности того или иного прочтения символов, подчинила в своей «речи» русские слова общим орфоэпическим и фонетическим законам чеченского языка. Так, ударение практически всегда падает на первый слог. Более того, в многосложных словах

система «редуцирует» некоторые гласные в соответствии с законами фонетики чеченского языка. Таким образом, некоторые заимствования из русского языка бывают порой трудноузнаваемы.

Плюсы прототипа системы синтеза чеченской речи.

1. Разработанная нами система в абсолютном большинстве правильно расставляет ударения в словах (за исключением поздних заимствований из русского языка). В целом, это и не должно было представлять особой сложности, поскольку ударение обычно падает на первый слог. Однако трудно было предусмотреть поведение системы по отношению к сложным словам, в которых приставки, а иногда и первая основа в словах с двумя или более корнями оказываются безударными. Тем не менее, обучение дало неплохой результат, и нейронная сеть обучилась расстановке ударений в чеченском языке без дополнительных алгоритмов.

2. В основу базы для обучения нейронной сети нами были положены различные тексты, содержащие словарный материал с абсолютно разным материальным оформлением. Например, в базе присутствуют разговорные тексты, где часто предложение состоит из одного короткого слова, и художественные произведения, изобилующие длинными, сложными предложениями со всевозможного рода паузами и интонациями; есть публицистические тексты, где много заимствованных слов из различных языков, и научные, в которых часто встречаются сложные слова, состоящие из нескольких морфем, вследствие чего в них представлено иногда по несколько редуцированных гласных и согласных, а также сложные, нетипичные для других текстов сочетания звуков и т. д. Таким образом, в нашей базе представлены всевозможные варианты звучания отдельных фонем, слогов, словосочетаний, пауз, интонаций. Все это позволило системе обучиться правильному чтению даже таких элементов, которые встречаются в чеченском языке не так часто.

3. В прототипе системы имеет значение качество голоса и его выразительность. Так, система не говорит «роботизированным» голосом, он звучит вполне по-человечески. Соблюдаются все интонации начала и конца фразы, паузы между словами естественные. В то же время нет излишних перепадов высоты звуков, неуместных интонаций. Такой результат, на наш взгляд, обусловлен тем, что при выборе диктора внимание обращалось не только на четкую дикцию и «чистоту» голоса. Диктор должен читать ровно, без перепадов, соблюдая интонацию, с умеренной выразительностью, но без эмоций.

4. При анализе различных программ синтеза речи для разных языков нам нередко приходилось наблюдать проблему скорости прочтения текста. Так, некоторые программы тянули гласные звуки, делая речь неестественно медленной, другие – выдавали поток звуков с такой скоростью, что порой трудно было с первого раза разобрать сказанное. Однако в нашем прототипе программы эта проблема полностью отсутствует.

5. Качество синтеза заметно улучшают дополнительные модули, работающие в программе. Один из важнейших – транскриптор, обрабатывающий входной текст до создания звукового файла. Четко прописанные алгоритмы правильного чтения той или иной буквы с указанием конкретных исключений и особенностей позволяют исключить вероятность многих ошибок, что было бы невозможно только одним обучением, основанным на частотности соответствия фонемы графеме.

6. Модуль, повышающий уровень качества синтезируемой программой речи, – так называемый нормализатор – это программа для расшифровки записей, не раскрытых побуквенно, она преобразует запись чисел в словесную форму. В рамках системы синтеза результат действия этой программы далее обрабатывается транскриптором, в результате чего синтезатор получает вместо цифр «расшифровку» латинскими символами. По такому же принципу осуществляется прочтение сокращений. Мы отказались от идеи полной расшифровки сокращенных записей, поскольку даже в естественной человеческой речи мы часто употребляем сокращенные названия. Однако прописали в нормализаторе алгоритмы их правильного прочтения: не «чэрэ», а «чээр» (ЧР), не «мэгэу», а «эмгэу» (МГУ) и т. д. Исключение было сделано для таких сокращений, которые действуют только на письме: и. д. кх. а, о. т. кх. а, м, см, г, к. Все эти сокращения, которые мы встречаем на письме, в устной естественной речи не используются. Поэтому мы прописали полное прочтение этих формулировок при наличии сокращений: «иштта дIа кхин а», «оцу тайпа кхин а», «метр», «кийла» и т. д.

7. Четкость произношения звуков можно назвать одним из плюсов синтезируемой речи. Несмотря на то, что в чеченском языке присутствуют фонемы, достаточно близкие по звучанию (четыре количественно различающихся варианта «а», гортанные звуки «I», «Iъ», «ъ», троичный ряд согласных – «б», «п», «пI»; «д», «т», «тI»; «г», «к», «кI» и т. д.), обучение нейронной сети позволило системе довольно четко различать произношение этих звуков, и, в целом, программа довольно четко разграничивает те же конечные «хь» и «х», граница между которыми в устной речи чеченцев в последние годы стала размываться.

Сегодня мы ведем доработку прототипа системы синтеза речи, начаты исследования в области устранения графической омонимии. Указанные нами минусы синтезатора, типичны и для других языков, в частности, для русского языка также проблемными являются слова-омографы, но здесь надо делать акцент на правильной расстановке ударений. В чеченском же языке проблема омографов обусловлена отсутствием специальных маркеров для дифтонгов «уо» и «иэ», а также долготы гласных.

Для решения этой проблемы существует три основных подхода, основанных на: (1) правилах; (2) статистике; (3) машинном обучении.

Мы решили остановиться на гибридном методе устранения графической омонимии, основанном на использовании машинного обучения и статистики. Для чеченского языка этот подход кажется нам наиболее подходящим, так как при чтении чеченских текстов читателю приходится анализировать контекст, чтобы правильно произнести слова-омографы.

Далее мы планируем создать алгоритм и программную реализацию нейронной сети, которая после обучения на подготовленной базе будет осуществлять классификацию омографов и их выявление в тексте [6]. Это позволит нам оптимизировать код созданной программы автоматического транскрибирования чеченских текстов и улучшить качество синтезируемой системой чеченской речи за счет распознавания в тексте омографов.

Для решения проблемы чтения поздних заимствований из русского языка или через русский, нами был проведен статистический частотный анализ текстов на выявление наиболее часто встречающихся заимствований из русского языка. После этого был создан текстовый файл, содержащий чуть более сотни слов из этой категории.

По устранении недостатков в работе прототипа синтезатора речи, мы предлагаем запустить повторное обучение системы синтеза чеченской речи.

* * *

Оценка качества полученного нами речевого сигнала на соответствие с международной оценкой качества синтезируемой речи MOS показала, что качество синтезируемой речи приближено к естественной и средний результат дал 4 балла. Разборчивость синтезируемой речи составила 85,2 %.

Анализ тестирования прототипа системы синтеза чеченской речи и его доработка открыла новые направления исследований, предварительный результат которых ориентировал нас на создание алгоритма и программной реализации нейронной сети, которая после обучения на подготовленной базе, будет осуществлять классификацию омографов и их выявление в тексте.

СПИСОК ЛИТЕРАТУРЫ

1. Израилова Э.С. «Фонетический алфавит» чеченского языка как основа системы синтеза речи // Научно-техническая информация. Сер. 2. – 2018. – № 2 – С. 35-39; Izrailova E.S. The Phonetic Alphabet of the Chechen Language as a Basis of a Speech-Synthesis System // Automatic Documentation and Mathematical Linguistics. – 2018. – Vol. 52, № 1. – P. 51–55. DOI: 10.3103/ S0005105518010077.
2. Израилова Э.С. Моделирование системы синтеза речи на основе глубоких свёрточных нейронных сетей // Известия КБНЦ РАН. – 2018. – №6(2).
3. Hideyuki Tachibana, Katsuya Uenoyama, Shunsuke Aihara. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. – URL: https://arxiv.org/pdf/1710.08969.pdf (дата обращения 09.12.2020)
4. Израилова Э.С. Особенности машинного обучения средствами CNN в рамках синтеза речи // Вестник ГГНТУ. Технические науки. – 2019. – Т. XV, № 2(16) – С. 29-35.
5. Соломенник А.И. и др. Оценка качества синтезированной речи: проблемы и решения // Изв. вузов. Приборостроение. – 2013. – Т. 56, № 2. – С. 38-42.
6. Израилова Э.С. Процесс создания системы синтеза чеченской речи // Известия РГПУ им. А. И. Герцена. – 2020. – №198.

Материал поступил в редакцию 18.02.21.

Сведения об авторах

ИЗРАИЛОВА Элиса Салаудиновна – старший научный сотрудник отдела прикладной семиотики Академии наук Чеченской Республики; старший преподаватель кафедры информатики и вычислительной техники Грозненского государственного нефтяного технического университета
364024, ЧР, г. Грозный, пр-кт им. М. Эсамбаева, 13
e-mail: uelisa@yandex.ru

БАДАЕВА Айшат Салауддиновна – научный сотрудник отдела прикладной семиотики Академии наук Чеченской Республики; заместитель директора по науке Института чеченского языка.
e-mail: ayshatbs@gmail.com

ДЛЯ ЗАМЕТОК