РОССИЙСКАЯ АКАДЕМИЯ НАУК

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
(ВИНИТИ)

НАУЧНО • ТЕХНИЧЕСКАЯ
ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 8 Москва 2013

ИНФОРМАЦИОННЫЕ ЯЗЫКИ

УДК 025.4.036:004.738.5

Г.Г. Белоногов, Р.С. Гиляревский, С.Н. Селетков, А.А. Хорошилов

О путях повышения качества поиска текстовой информации в системе Интернет

Утверждается, что в системе Интернет качество поиска текстовой информации низкое. Причиной этого являются сложившиеся веками неправильные представления о смысловой структуре текстов. Предлагаются методы повышения качества поиска в системе Интернет.

Ключевые слова: *концептуальная структура текстов, поиск текстовой информации, качество поиска*

ВВЕДЕНИЕ

Первые универсальные цифровые электронные вычислительные машины появились в США и Англии в конце 40-х годов прошлого столетия (машины типа ЭНИАК и ЭДСАК), а первые советские машины подобного рода (машина «Стрела») – в начале 50-х годов. Вскоре стало ясно, что это не только вычислительные машины, но что они могут решать и более широкий класс задач, например, таких как логический вывод, распознавание образов, хранение и поиск информации, машинный перевод текстов с одних естественных языков на другие.

Электронные вычислительные машины стали использоваться, прежде всего, в военных ведомствах, где большое значение придавалось надежности их функционирования. Поэтому почти с самого начала делались попытки организовать одновременную взаимосвязанную («параллельную») работу нескольких машин. В 1969 г. в США была создана первая компьютерная сеть, которая связала четыре научных учреждения, а в 1973 г. появилась и международная сеть.

В СССР в качестве инициатора создания общегосударственной сети вычислительных центров, предназначенной для решения задач обороны страны и управления народным хозяйством, выступил Анатолий Иванович Китов – заместитель начальника ЦНИИ 27 МО по НИР. В начале 60-х годов прошлого столетия он подготовил свои предложения по этому вопросу и оформил их в виде отдельной книги. В книге наряду с позитивными положениями содержалась также критика руководства Министерства Обо-

роны СССР за недооценку важности электронной вычислительной техники как средства повышения эффективности управления войсками.

А.И. Китов направил свои предложения в ЦК КПСС. Но они были переадресованы Министерству Обороны СССР. А руководство этого министерства встретило их враждебно и отвергло. Более того, там решили примерно наказать автора предложений: А.И. Китов был снят с занимаемой должности, уволен из армии и исключен из партии. Так советское государство расправилось с одним из своих выдающихся ученых. Но, к чести нашей страны, по прошествии нескольких десятилетий профессор А.И. Китов был реабилитирован и признан пионером отечественной кибернетики и информатики.

Международная компьютерная сеть быстро развивалась. В 1980-х годах она получила широкое признание и стала называться системой "Интернет" (международная сеть). Эта огромная компьютерная сеть стала объединять тысячи меньших сетей, разбросанных по всему миру.

В конце 1980-х – начале 1990-х гг. появилась новая технология, ставшая вскоре главной и всеобъемлющей в системе Интернет – это World Wide Web (всемирная паутина), кратко – WWW. Ее автор английский ученый Тим Бернес-Ли и его соратники создали средства, позволяющие связывать информацию из различных источников и делать ее доступной из любой точки сети [1]. Эти средства включают унифицированный идентификатор ресурсов (Uniform Resource Locator – URL), язык разметки гипертекста (HyperText Mark-Up Language - HTML), протокол передачи гипертекста (HyperText Transfer Protocol – HTTP).

В основу структуры "всемирной паутины" была положена концепция *гипертекста*, предложенная Т. Нельсоном. Понятие *гипертекст* различные авторы характеризуют по-разному. Например, в книге [1] на стр. 40 это понятие характеризуется как способ хранения и манипулирования информацией, при котором она представлена в виде сети связанных между собой узлов. Каждый узел может содержать текст, графику, видео- или аудиоинформацию: Доступ к узлам - их просмотр или манипулирование ими – может осуществляться в интерактивном режиме. Отдельный документ именуется в системе WWW ***"веб-страница"***, а множество документов, принадлежащих отдельному лицу или организации, ***- "сайт"*** (site – место).

Обмен информацией между компьютерами организуется согласно определенному протоколу программами двух типов: программами-серверами и программами-клиентами. Программа-сервер обеспечивает хранение информационных ресурсов и выдачу их по запросам программ-клиентов (соответственно, компьютер, где размещаются ресурсы, тоже называют сервером). Программа-клиент формирует запросы к серверу, принимает и интерпретирует для пользователя получаемую от сервера информацию. Программы-клиенты на компьютере пользователя в системе WWW получили название браузеров (от английского глагола browse – просматривать, пролистывать).

Поиск документов в текстовых базах данных ведется по их формализованным описаниям – по так называемым индексам. **Индексы** представляют собой инверсные файлы, состоящие из перечней различных слов, входящих в состав документов (за исключением так называемых "стоп-слов"), с указанием всех адресов их вхождения в документы. Автоматическое индексирование документов осуществляется специальными программами-роботами, которые в различных поисковых системах называются по-разному: spider (паук), crawler (ползун), worm (червь).

Система Интернет – это выдающееся достижение человечества. Она имеет шансы превратиться в глобальную информационную модель мира, а в дальнейшем - и в ноосферу (сферу человеческого разума).

В настоящее время в этой системе наиболее совершенными являются средства вычислительной техники и средства связи. Значительно скромнее выглядят средства смысловой обработкой текстовой информации, в частности, средства ее поиска. И здесь, к сожалению, нельзя не согласиться с авторами статьи [2], которые пишут: *"Сегодня пользователи глобальной сети вынуждены констатировать тот факт, что сравнительно часто им приходится сталкиваться с ситуацией, когда* ***количество полезной информации переходит в разряд бесконечно исчезающей величины на фоне информационного шума.***

В этих условиях перед разработчиками информационно-поисковых систем ставятся вопросы о необходимости создания новых программных средств, которые основываются на новых принципах и подходах к информационному поиску, включающих анализ смысла искомых объектов или запросов (семантический анализ)."

Есть и еще одно прямо-таки обескураживающее обстоятельство, имеющее место в системе Интернет. Оказывается, **в этой системе в настоящее время хранится уже более одного триллиона веб-страниц, и поиск в этом массиве ведут более 2000 поисковых систем** [3]. **Но самая мощная из них – система Google –охватывает только 4,5% этого массива, а остальные – существенно меньше, как правило, менее одного процента.** Так что задача радикального улучшения качества поиска информации в системе Интернет является весьма актуальной.

Главная причина плохого качества поиска информации в системе Интернет заключается в том, что ее разработчики исходят из неправильных представлений о смысловой структуре текстов. Они полагают, что основным средством обозначения понятий в текстах являются отдельные слова, а не фразеологические словосочетания. Такая точка зрения господствовала веками и, в основном, является доминирующей и поныне. Это тормозит исследования и разработки, связанные с автоматической смысловой обработкой текстовой информации.

Та истина, что устойчивые фразеологические словосочетания являются основным средством выражения наименований понятий в текстах и что они используются в такой роли в сотни раз чаще, чем отдельные слова, была впервые установлена в ЦНИИ 27 МО СССР и в ВИНИТИ АН СССР в процессе масштабных статистических исследований современных русских и английских текстов, прово-

дивших в течение ряда десятилетий. Эта истина была многократно подтверждена в процессе разработки и эксплуатации ряда систем автоматической обработки текстовой информации – систем автоматического индексирования документов, их автоматической классификации и поиска, систем автоматического перевода текстов с одних естественных языков на другие.

Следует заметить, что в Советском Союзе первыми научно-исследовательскими организациями, которые начали вести работы по созданию автоматизированных информационных систем, были ЦНИИ 27 МО и ВИНИТИ АН СССР [4-10]. Именно эти организации долгое время были лидерами в системе Министерства Обороны и в Государственной Системе Научно-Технической Информации (ГАСНТИ). Именно в этих организациях впервые сложились правильные представления о смысловой структуре текстов [8-15]. Их опыт был бы полезен при решении проблем, ныне стоящих перед системой Интернет.

СМЫСЛОВАЯ СТРУКТУРА ТЕКСТОВ

Как известно, естественный язык является универсальным средством общения между людьми – средством восприятия, накопления, хранения, поиска и передачи информации. Более того, он является инструментом мышления человека [16-19]. Психологи считают, что естественный язык представляет собой вторую сигнальную систему человека, функционирующую на основе первой сигнальной системы, которая, в свою очередь, работает как система врожденных безусловных рефлексов, возникающих под воздействием сигналов, получаемых от зрительных, слуховых, тактильных и других рецепторов [17]. Языковые сигналы лишь инициируют мыслительные процессы, происходящие во внутреннем духовном мире человека, в его “душе”, но не определяют их полностью. По мнению психологов, интерпретация речевых сигналов человеком (их понимание) происходит с учетом опыта, накопленного им в течение всей своей жизни.

В процессе общения людей друг с другом одни и те же явления природы и общества могут описываться с различной степенью общности и с помощью различных языковых средств И это осложняет автоматический смысловой анализ текстов. Но есть еще одно явление, усугубляющее положение. Это – существование наряду с видимым текстом еще и **подтекста,** подразумеваемого текста. Подчеркивая важность этого явления, российский ученый В.А. Звегинцев предлагал наряду с грамматикой видимого текста разрабатывать также и *“грамматику подтекста*”. Для решения проблемы подтекста лингвисты пытались ввести в научный обиход так называемые **пресуппозиции** – предложения, характеризующие подтекст [20]. Однако из этого ничего не вышло. Попытка оказалась неудачной.

Радикальным решением здесь могло бы быть построение модели “души” человека - модели его внутреннего мира. Но эта задача очень сложная и над ней, как и над проблемой “искусственного интеллекта”, довлеет “проклятие размерности”. Она будет решена не скоро (если вообще когда-либо будет решена). Однако это не значит, что в настоящее время здесь не надо искать более простых решений.

В этой связи можно провести некоторую аналогию с космонавтикой. Известно, что голубая мечта космонавтики – это дальние космические полеты. Но если в настоящее время такие полеты невозможны, то это не означает, что надо отказаться от воздушных полетов или от полетов в ближнем космосе. Так и в рассматриваемой ситуации: если пока нет возможности найти точное решение проблемы, то можно ограничиться и приближенным решением, если оно существенно лучше, чем существующее.

Естественный язык рассматривается в *лингвистике* как некоторая *знаковая система*. По мнению Ф. де Соссюра – одного из основоположников современной науки лингвистики и науки семиотики - *языковые знаки состоят из двух компонент: из **означающего** и **означаемого*** [18]. ***Означающее*** – *это звуковой или графический образ знака, а **означаемое** – соответствующее ему **понятие**.*

Понятие - это социально значимый мыслительный образ, за которым в языке закреплено его наименование в виде отдельного слова или, значительно чаще, в виде устойчивого фразеологического словосочетания. Под устойчивыми фразеологическими словосочетаниями мы будем понимать не только идиоматические выражения и терминологические словосочетания, но и любые повторяющиеся отрезки связных текстов длиной от двух до десяти-пятнадцати слов (более длинные устойчивые словосочетания встречаются редко).

В развитых языках мира (русском, английском, немецком, французском, испанском и др.) количество различных наименований понятий достигает нескольких сотен миллионов. Большинство из них обозначаются словосочетаниями, смысл которых не сводим к смыслу составляющих их слов.

По современным представлениям *наиболее устойчивыми единицами смысла являются **понятия***. Они занимают центральное место в языке и речи и являются теми базовыми строительными блоками, на основе которых формируются смысловые единицы более высоких уровней. *Второй по значимости единицей смысла является **предложение** (высказывание). Из предложений формируются различного рода **сверхфразовые единства***, которые представляются в виде последовательностей предложений (связного текста).

*Основной чертой предложений является их **предикативность*** – т. е. то их свойство, что в них утверждается наличие у объектов определенных признаков и отношений [8, 20]. Свойством предикативности обладают и высказывания, формулируемые на формализованных языках

Смысл текстовых документов выражается с помощью единиц смысла, входящих в их состав. Как уже было указано, в естественных языках базовыми единицами смысла являются понятия. Поэтому **центральной процедурой любых систем автоматической смысловой обработки текстов должна быть процедура их семантико-синтаксического концептуального (понятийного) анализа. По нашему мнению, она должна быть реализована прежде**

всего как процедура фразеологического концептуального анализа на основе мощных словарей наименований понятий. При этом следует опираться на адекватные семантико-синтаксические модели текстов, в которых понятия представляются преимущественно фразеологическими словосочетаниями.

Как уже указывалось, обычно в разных текстах одни и те же объекты и процессы могут описываться с различной степенью общности и с помощью различных языковых средств. Поэтому при решении задач автоматической смысловой обработки текстовой информации необходимо в той или иной мере учитывать такие явления как **словоизменение, синонимия, гипонимия** (родо-видовые отношения), **разнообразие средств выражения межфразовых связей.**

Явление **словоизменения** может быть учтено путем применения процедур автоматического морфологического анализа слов и отождествления различных форм слов по их основам. Для учета явлений **синонимии** и **гипонимии** необходимо использовать **словарь синонимов, гипонимов** (более узких по объему понятий) **и гиперонимов** (более широких по объему понятий) как на уровне отдельных слов, так и на уровне фразеологических словосочетаний.

Сложнее дело обстоит с учетом таких явлений как **вариация обозначений одних и тех же понятий в связных текстах.** В этих текстах наименование понятия, выраженное фразеологическим словосочетанием, может быть сначала представлено в своей исходной форме, а затем, в последующих предложениях, - в сокращенных вариантах. Оно может быть также заменено на наименование родового понятия или на местоимение. Это имеет место, например, в следующем отрезке текста: “По дороге бежала ***большая рыжая лохматая собака***. ***Собака*** прихрамывала на левую переднюю лапу. ***Животное*** куда-то спешило. ***Оно*** было явно чем-то обеспокоено”.

СОСТАВЛЕНИЕ СЛОВАРЕЙ НАИМЕНОВАНИЙ ПОНЯТИЙ

Как уже было сказано, понятия являются базовыми единицами смысла. Поэтому составление словарей наименований понятий должно быть одной из центральных задач при построении автоматизированных информационно-поисковых систем. До середины прошлого века словари составлялись вручную. В них обычно не приводилось никаких сведений о частоте использования лексических единиц в текстах. Такие сведения стали указываться только в частотных словарях, которые начали составляться в конце 50-х годов прошлого века в связи с развертыванием работ по машинному переводу и информационному поиску. Составление частотных словарей “вошло в моду” и этим стали заниматься многие коллективы.

Первый частотный словарь русского языка был составлен американским ученым Джоссельсоном (Josselson) по текстам протяженностью около *полумиллиона* слов. Для этого он использовал общественно-политические и художественные тексты, изданные в 19-20 веках. Несколько позднее, **в 1958 г., одним из авторов настоящей статьи** *(Г.Г. Белоноговым)* **был составлен “Частотный словарь современных русских оперативно-тактических текстов”**. Объем обработанных текстов был примерно такой же, что и у Джоссельсона, но они имели более узкую тематику и относились к более короткому интервалу времени (1957-1958 гг.). Словарь составлялся в ЦНИИ 27 МО СССР.

Начатые в ЦНИИ 27 МО работы по составлению словарей были далее продолжены в ВИНИТИ АН СССР. Так **в 1985 г. в этом институте был обработан корпус политематических текстов общим объемом более 70 млн слов (в книжном представлении это было бы 560 книг по 400 страниц каждая). По этому массиву текстов был составлен грамматический словарь современного русского языка объемом около 180 тыс. лексических единиц.**

В ВИНИТИ проводилась также работа по составлению частотных словарей ключевых слов и словосочетаний, содержащихся в поисковых образах документов, представленных в базах данных. В этом институте функционировала мощная система автоматизированной подготовки и издания реферативных журналов по широкому спектру областей науки и техники. При этом в течение года обрабатывалось более одного миллиона документов на различных языках мира.

По каждому документу составлялись его библиографическое описание, реферат и поисковый образ в виде набора ключевых слов и словосочетаний, характеризующих основное смысловое содержание этого документа. Поисковые образы документов были предназначены, прежде всего, для составления предметных указателей к реферативным журналам и могли содержать различное число ключевых слов и словосочетаний. Обычно оно не превышало 10-12 наименований понятий и в среднем было равно пяти.

В ВИНИТИ была разработана инструкция по составлению поисковых образов документов, в которой рекомендовалось применять для описания документов преимущественно однословные наименования понятий. Однако при массовом индексировании документов сотрудники научно-отраслевых отделов и внештатные референты ВИНИТИ игнорировали эту инструкцию и описывали документы в терминах, принятых в соответствующих областях знаний (а это были по большей части фразеологические словосочетания). В результате **был создан и зафиксирован на машиночитаемых носителях богатейший фонд наименований понятий современной науки и техники**.

Для составления частотных словарей ключевых слов и словосочетаний был взят пятилетний массив поисковых образов документов (1983-1987 гг.). Этот массив включал в свой состав более пяти миллионов поисковых образов, содержащих более 25-ти млн наименований понятий (в книжной форме представления этот массив состоял бы из 400 книг объемом 400 страниц каждая).

По исходному массиву поисковых образов документов сначала составлялись отраслевые частотные словари ключевых слов и словосочетаний, затем **эти словари были сведены в один политематический частотный словарь. Он получился объемом более одного миллиона наименований понятий.**

Далее, в течение 80-х и 90-х гг. прошлого столетия, в ВИНИТИ постоянно велась масштабная работа по составлению машинных словарей различного назначения. В 1990-х годах она была связана преимущественно с созданием и развитием системы фразеологического машинного перевода текстов с русского языка на английский и с английского на русский - системы RETRANS (Russian-English TRANslation System) [12].

Промышленная версия этой системы была создана в 1993 г. и сразу стала использоваться в России (в ВИНИТИ РАН, ВНТИЦентре и Миннауки) а также в правительственных организациях США (Госдепартамент) и Франции (агентство CEDOCAR). Госдепартамент США финансировал исследования и разработки ВИНИТИ по системе RETRANS в течении двух лет (1994-1995 г.г.), агентство CEDOCAR – в течение шести лет (1994-1999 г.г.). При этом в течение четырех лет для агентства CEDOCAR с помощью системы RETRANS переводились с русского языка на английский материалы из реферативного журнала ВИНИТИ.

К 2003 году политематический машинный словарь системы RETRANS имел объем уже более двух с половиной миллионов словарных статей в одном направлении перевода (в книжном представлении это было бы примерно 50 томов по 1000 страниц). Большинство входов в этот словарь (80%) являлись фразеологическими словосочетаниями длиной от 2-х до 15-ти слов.

Кроме основного политематического словаря для системы RETRANS было составлено еще более одного десятка дополнительных тематических словарей суммарным объемом более 250 тыс. словарных статей (примерно 5 томов по 1000 страниц). В этих словарях содержались только приоритетные для каждой тематики переводные эквиваленты наименований понятий, отличные от приоритетных переводных эквивалентов политематического словаря.

Машинные словари системы RETRANS постоянно пополнялись как в интерактивном режиме (в процессе перевода текстов с одного языка на другой), так и путем автоматического составления словарей по двуязычным текстам (кстати, это делалось впервые в мире!). Последняя операция выполнялась неоднократно. Так, **в 2010 г. были составлены англо-русский и русско-английский политематические фразеоло-гические словари по массиву двуязычных (русских и английских) заголовков документов объемом более двух миллионов пар заголовков** [24]. Пары двуязычных заголовков были извлечены из реферативных баз данных ВИНИТИ. **Составленные словари (англо-русский и русско-английский) получились объемом более одного миллиона словарных статей каждый.**

В табл. 1 приведено распределение длин наименований понятий в русском словнике русско-английского политематического словаря системы RETRANS.. В этом словнике длина наименований понятий варьирует в пределах от одного до пятнадцати слов, а самыми частыми словосочетаниями являются двухсловные и трехсловные. Средняя длина словосочетаний в словнике оказалась равной 2,9 слова.

Таблица 1

Распределение длин русских наименований понятий в русско-английском словаре системы RETRANS

Кол-во слов в наименовании	*Относительная частота*	*Кол-во слов в наименовании*	*Относительная частота*
1	0,136	9	0,004
2	0,384	10	0,002
3	0,218	11	0,001
4	0,127	12	0,0005
5	0,068	13	0,0004
6	0,035	14	0,0002
7	0,016	15	0,0001
8	0,008		

Как уже указывалось, в различных речевых актах одни и те же понятия могут иметь разные наименования (могут описываться различными сочетаниями слов). Это иллюстрируется табл. 2., в которой представлен краткий перечень пар русских синонимических словосочетаний.

Таблица 2

Русские синонимические словосочетания

1. абсолютная жесткость / бесконечно большая жесткость;
2. абсолютная слепота / полная слепота;
3. абсолютная температура / температура Кельвина;
4. базовый список / основной перечень;
5. вакуумное напыление / вакуумное осаждение;
6. кривая второго порядка / коническое сечение;
7. гармонический синтезатор / синтезатор Фурье;
8. как следствие / по этой причине;
9. неожиданно / ни с того, ни с сего;
10. наклонный путь для сортировки вагонов / путь сортировочной горки;
11. наклоны головы в поперечной плоскости / наклоны головы к правому и левому плечам;
12. сильное раскаяние / угрызения совести.

Как можно увидеть из табл. 2, наименования понятий-синонимов обозначаются различными словосочетаниями. Причем в п.п. 4, 6, 8, 9 и 12 словосочетания, стоящие слева и справа от косой черты, одинаковых слов совсем не содержат.

Более представительный фрагмент словаря синонимических отношений между фразеологическими словосочетаниями приведен в *Приложении* к настоящей статье. Этот словарь был сформирован на основе англо-русского политематического фразеологического словаря системы RETRANS и содержит более 290 тыс. синонимических рядов русских словосочетаний (рядов переводных эквивалентов английских наименований понятий). Словарь имеет объем 19 мегабайт, что в книжном представлении составило бы 19 книг объемом по 400 страниц.

Автоматическое составление фразеологических словарей наименований понятий по текстам – более

трудная задача, чем задача составления словарей слов, так как, в отличие от слов, границы фразеологических словосочетаний в текстах никак не обозначены (они "отмечены" только в сознании человека).

Фразеологические словари могут составляться по текстам двумя способами: - с контролем по тезаурусу и без такого контроля. В первом случае для выделения наименований понятий из текстов необходимо иметь большой базовый политематический словарь-тезаурус, обеспечивающий хорошее покрытие текстов (не менее 99%). Во втором случае базовый словарь не используется, а фразеологические словосочетания выделяются из текстов статистическими методами.

В обоих случаях составление словаря по тексту начинается с его концептуального анализа (его членения на наименования понятий). Далее, по массиву выделенных наименований понятий составляется частотный словарь. Наконец из составленного словаря исключаются низкочастотные наименования понятий (например, наименования с частотой $f<3$).

Метод составления словарей с контролем по тезаурусу дает более точные результаты, чем метод их составления без контроля по тезаурусу,. но он не позволяет выявлять "новые" наименования. понятий. Второй метод это позволяет. Но здесь возможны ошибки. Чтобы их избежать, нужно использовать преимущественно высокочастотную часть словаря и проводить его постредактирование.

Когда речь идет о составлении словарей с контролем по тезаурусу, то возникает естественный вопрос: а откуда брать базовый политематический словарь-тезаурус? Ответ на этот вопрос по существу дан выше. Ведь мы уже говорили о том, что в ВИНИТИ РАН в течение десятков лет велись масштабные работы по автоматическому составлению словарей, в результате которых были созданы русские, английские, русско-английские и англо-русские словари наименований понятий большого объема. На основе этих словарей можно за короткие сроки (за несколько недель!) создавать нужные политематические словари (и английские, и русские).

Словари, предназначенные для автоматической обработки текстов, должны содержать информацию о внеконтекстных ассоциативных смысловых связях между понятиями, как минимум, о связях типа "синонимия" и "гипонимия" (родо-видовые связи). Выявление таких связей – более трудная задача, чем составление словарей. Но они совершенно необходимы для распознавания смысловой близости текстов.

В 27 ЦНИИ МО в конце 70-х годов прошлого столетия в связи с разработкой автоматизированных информационно-поисковых систем была начата работа по составлению словаря синонимов, гипонимов и гиперонимов русских слов. Он получился объемом около 25.000 словарных статей. Затем работа над словарем была продолжена в ВИНИТИ РАН. При этом была использована информация, содержащаяся в семидесяти тезаурусах системы ГАСНТИ (Государственной Автоматизированной Системы Научно-Технической Информации). Словарь возрос до объема 32.000 словарных статей. В табл. 3 приведен фрагмент словаря синонимов и гиперонимов.

Таблица 3

Фрагмент словаря синонимов и гиперонимов

осветитель / светильник / светоизлучатель / прибор / аппаратура / устройство
освещенность / освещение / свет / характеристика
освидетельствование / изучение / обследование / действие / процесс
освидетельствовать / обследовать / исследовать
обрабатывание / обработка / процесс
освоение / применение / использование / употребление
осетины / нация / национальность / народ / население
осетр / рыба / организм
осилить / преодолеть / одолеть / победить
осина / дерево / растение
ослабление / снижение / спад / убавление / уменьшение
осмий / металл / элемент / проводник / вещество
оснастка / оснащенность / экипировка / оснащение / оборудование
оснащать / оснастить / обеспечить / снабдить
оснащение / оснащенность / экипировка / оснастка / снабжение / обеспечение
оснащенность / экипировка / оснащение / оснастка
основательный / веский / убедительный

С помощью этого словаря в ВИНИТИ был проведен эксперимент по "избыточному индексированию" рефератов документов по 16-ти различным тематикам. В процессе эксперимента слова из текстов рефератов дополнялись их синонимами и гиперонимами. Оказалось, что по всем тематикам количество слов в рефератах в среднем утроилось, что свидетельствует о высоком качестве словаря**.**

При создании методов автоматической обработки информации в системе Интернет необходимо в максимальной степени учитывать лексическое богатство современных естественных языков. А оно почти необъятно. Более того, в связи с процессами, происходящими в жизни общества, понятийный и, соответственно, лексический состав языков постоянно меняется. Если бы даже в какой-то момент и удалось его зафиксировать, то через некоторое время он бы изменился.

Выявить полностью лексический состав таких языков как, например, русский, английский, французский или испанский, довольно трудно (объем словаря наименований понятий каждого из этих языков оценивается примерно в 300-350 млн словарных статей). Но выявить статистическое ядро лексики этих языков можно за относительно короткое время. Оно будет иметь объем примерно 10-15 млн словарных статей и будет обеспечивать покрытие политематических текстов, как минимум, на 99,99%. Кроме того, для наиболее актуальных тематических областей можно составить дополнительные тематические словари объемом примерно в полмиллиона словарных статей. Тематические словари должны использоваться на фоне политематического словаря как его дополнение.

Составление словарей и их ведение (пополнение, корректировка) – нелегкая и трудоемкая задача. Для ее решения необходимо создавать специальную *ав-*

томатизированную словарную службу. Эта служба должна опираться на развитую систему программных средств, включающую процедуры морфологического, синтаксического и концептуального анализа и синтеза текстов, процедуры составления словарей и процедуры, необходимые для выполнения операций над словарями.

Самой трудной задачей в системе автоматизированной словарной службы является установление внеконтекстных устойчивых ассоциативных смысловых связей между понятиями (так называемых **парадигматических** связей). Наиболее важными из них являются связи типа **синонимия** и **гипонимия** (родо-видовые связи).

Мы уже указывали, что у авторов настоящей статьи есть определенные успехи в решении этой задачи. Но этого недостаточно. Чтобы достичь большего, необходимо и далее продолжать эту работу с широкой опорой на средства автоматизации. Основные направления такой работы могут быть следующие:

– установление парадигматических связей между фразеологическими словосочетаниями по их словарному составу с использованием словаря смысловых связей слов (парадигматических связей между отдельными словами);

– составление двуязычных словарей по двуязычным текстам большого объема и объединение различных вариантов перевода одних и тех же фразеологических словосочетаний в синонимические ряды;

– использование информации о парадигматических связях межу понятиями, содержащуюся в ранее составленных тезаурусах и классификаторах.

ПУТИ ПОВЫШЕНИЯ КАЧЕСТВА ПОИСКА ТЕКСТОВОЙ ИНФОРМАЦИИ В СИСТЕМЕ ИНТЕРНЕТ

В настоящее время в системе Интернет для поиска текстов используются, как правило, три типа их формализованных описаний: гипертексты, инверсные файлы и системы классификации. **В этих формализованных описаниях в качестве основных единиц смысла выступают либо отдельные слова (в инверсных файлах) либо сверхфразовые единства (в гипертекстах и в системах классификации). Но, к сожалению, главная форма выражения наименований понятий - устойчивые фразеологические словосочетания – не используется или почти не используется.**

Между тем, введение в формализованные описания текстов фразеологических словосочетаний позволило бы решить в системе Интернет сразу две важные задачи: 1) повысить полноту поиска информации, 2) повысить его точность. Полноту поиска можно было бы повысить за счет учета явлений синонимии и гипонимии словосочетаний как целостных единиц, а точность – за счет того, что в словосочетаниях в основном устраняется многозначность слов и связанные с ней ошибки.

Ориентация на фразеологические словосочетания как на основную форму представления наименований понятий в естественных языках позволила бы более точно учитывать семантико-синтаксическую структуру текстов и построить более эффективную систему поиска информации. Такая система должна включать в свой состав следующие компоненты:

1. Мощный политематический словарь наименований понятий объемом не менее 3 млн словарных статей, состоящий преимущественно из фразеологических словосочетаний. В словаре должны содержаться сведения об отношениях синонимии и о родовидовых отношениях между понятиями.

2. Дополнительные тематические словари наименований понятий для приоритетных тематических областей.

3. Концептуальные образы документов (КОДы), содержащие наиболее значимые наименования понятий, адекватно отражающие содержание этих документов.

4. Рефераты документов, включающие в свой состав их заголовки и наиболее информативные предложения.

5. Программные средства, необходимые для автоматического индексирования, автоматического реферирования и поиска документов.

Автоматическое индексирование документов (построение КОДов) должно осуществляться на основе их концептуального анализа. В состав КОДов должны включаться наиболее информативные наименования понятий. Информативность наименований понятий оценивается их “весом” P, который определяется как. произведение .третьей степени длины наименования понятия, измеряемой количеством слов в нем, и частоты встречаемости этого наименования в тексте:

$$P = f\, n^3,$$

где f – частота встречаемости наименования понятия в тексте; n – количество слов в этом наименовании.

Но при этом следует учитывать, что однословные наименования понятий повторяются в текстах значительно чаще, чем словосочетания, хотя они бывают, как правило, менее информативными. Поэтому нельзя допускать, чтобы числовые оценки “весов” однословных наименований понятий были равны или, тем более, превосходили третью степень количества слов в двухсловных словосочетаниях (число 8). По этой причине всем однословным наименованиям понятий с частотой $f > 7$ нужно присваивать частоту 7. После присвоения “весов” наименованиям понятий они сортируются по их убыванию, и в составе поискового образа документа оставляются только наиболее значимые наименования понятий.

Автоматическое реферирование документов необходимо для быстрой оценки результатов поиска (полные тексты документов для этого непригодны). В рефератах должны быть представлены заголовки документов и несколько их наиболее значимых предложений. Значимость предложений определяется как сумма “весов” наименований понятий, входящих в их состав.

Поисковые запросы должны формулироваться на естественных языках. В процессе поиска они должны подвергаться автоматическому концептуальному анализу и по его результатам должны строиться их концептуальные образы. Далее наименования поня-

тий этих концептуальных образов должны дополняться их синонимами, гипонимами и гиперонимами.

В процессе поиска обогащенный концептуальный образ запроса сравнивается с концептуальными образами документов поискового массива. При этом для каждой пары сравниваемых концептуальных образов определяется коэффициент их смысловой близости. Этот коэффициент равен отношению суммы весов исходных наименований понятий запроса, связанных по смыслу с наименованиями понятий концептуального образа документа, к сумме весов всех исходных наименований понятий запроса. Исходное наименование понятия запроса считается связанным по смыслу с одним из наименований понятия документа, если эти наименования совпадают, или если наименование понятия КОДа совпадает с одним из синонимов, гипонимов или гиперонимов исходного понятия запроса.

По окончании процесса сопоставления всех наименований понятий концептуальных образов запроса и документа определяется коэффициент смысловой близости запроса и документа. Документ считается релевантным, если коэффициент его смысловой близости запросу превышает заданный порог значимости.

После обработки всех документов поискового массива документы, связанные по смыслу с запросом, упорядочиваются по убыванию значений коэффициентов их смысловой близости запросу и пользователю выдаются рефераты наиболее значимых документов. Этот результат может быть представлен в виде нескольких эшелонов выдачи, содержащих информацию о документах с различной степенью их смысловой близости запросу.

При поиске информации в системе Интернет возникает необходимость обращаться к разноязычным базам данных. В этой связи становится актуальной задача машинного перевода текстов с одних естественных языков на другие. Проблемой машинного перевода текстов человечество занимается с середины прошлого века, но до начала 1990-х гг. успехи в ее решении были весьма скромными.

Дело в том, что разработчики систем машинного перевода ориентировались на семантико-синтаксический преимущественно пословный перевод текстов, а это – тупиковое направление, так как наименования одних и тех же понятий в разных языках, как правило, не являются пословными переводами друг друга. Выход из этого тупика может быть только один – отказаться от преимущественно пословного перевода текстов (по “значениям” слов), заменив его на преимущественно фразеологический перевод – перевод по наименованиям понятий, выраженным устойчивыми фразеологическими словосочетаниями. Концепция фразеологического машинного перевода была разработана в ЦНИИ 27 МО в 70-х годах прошлого столетия, а в 1993 г. в ВИНИТИ РАН была создана первая промышленная версия такого перевода. В дальнейшем было создано еще несколько версий этой системы [8, 11, 12, 15, 23, 24].

При поиске в разноязычных базах данных требуется переводить поисковые запросы на языки поисковых массивов, и результаты поиска - на языки запросов. Для этого необходимо иметь мощную систему машинного перевода текстов со многих естественных языков на многие другие. Известно, в мире существует около 2500 различных языков, а число их парных сочетаний превосходит 3 млн. Но, если бы даже различных языков было не более сотни, то и тогда для перевода текстов с любого языка на любой другой потребовалось бы около пяти тысяч систем перевода. А это трудновыполнимая задача.

Выходом из этого затруднения мог бы быть отказ от построения систем машинного перевода с любого языка на любой другой, и вместо этого осуществлять перевод с помощью *языка-посредника*. Тогда можно было бы существенно сократить число разрабатываемых систем перевода. Так, например, в случае ста различных языков вместо 4.950 пришлось бы создавать только 99 систем перевода (в пятьдесят раз меньше!).

Идея языка-посредника была высказана еще на рубеже конца 1950-х и начала 1960-х гг. прошлого столетия. Но она тогда не была реализована, так как для этого не было необходимых условий. Однако в настоящее время, в связи с улучшением качества машинного перевода, к этой идее можно было бы вернуться.

Среди различных предложений по языку-посреднику, выдвинутых пионерами машинного перевода, было предложение использовать в качестве такого языка искусственный язык Esperanto. На наш взгляд это неразумно, так как любой искусственный язык, имеет более бедную систему понятий, чем естественные языки, и не годится в качестве языка-посредника. В таком качестве может выступать только один из естественных языков с достаточно богатой системой понятий (например, русский, английский, немецкий или французский).

Скорее всего, развитие машинного перевода пойдет по пути разработки двуязычных систем перевода в интересах наиболее развитых стран мира. А по мере их создания постепенно будет появляться возможность перевода текстов и между парами языков, не обеспеченными изначально системами перевода, через посредство имеющихся в наличии систем. И, возможно, только на более позднем этапе развития будет достигнуто соглашение о едином языке-посреднике или о нескольких таких языках.

ЗАКЛЮЧЕНИЕ

Подводя итоги нашим рассуждениям о путях решения задачи повышения качества поиска текстовой информации в системе Интернет, мы хотели бы подчеркнуть важность первоочередного решения следующих проблем:

Проблема первая – выявление понятийного и фразеологического богатства естественных языков. Это важно потому, что понятия являются базовыми и наиболее устойчивыми единицами смысла и в мышлении, и в языке, и в речи.

Проблема вторая – выявление наиболее устойчивых внеконтекстных ассоциативных смысловых связей между понятиями. Эта проблема по существу

является частью первой, так как речь здесь идет об описании смыслового содержания понятий, а оно наиболее полно раскрывается в системе их ассоциативных связей друг с другом.

Проблема третья – разработка базовых процедур семантико-синтаксического анализа и синтеза текстов на основе их фразеологического и концептуального анализа и синтеза.

Проблема четвертая – разработка программных средств поиска в полнотекстовых базах данных, в которых в качестве основной формы выражения наименований понятий рассматриваются не отдельные слова, а устойчивые фразеологические словосочетания.

Перечисленные проблемы являются наиболее приоритетными, так как они возникают при решении любых достаточно сложных задач автоматической смысловой обработки текстовой информации.

СПИСОК ЛИТЕРАТУРЫ

1. Захаров В.П. Информационные системы (документальный поиск). – Санкт-Петербургский государственный университет, 2002
2. Мельников В.О., Максимов О.А., Меликян Г.С. Характеристика информационно-поисковых систем Интернет: теоретические и практические аспекты // Научно-техническая информация. Сер. 2. – 2009. – № 2. – С. 15-23.
3. Селетков С.Н. Теоретические проблемы информатики. Том 3. Проблемы. эффективности использования мировых информационных ресурсов / под общей ред. К.И. Курбакова.– М.: РЭА им. Г.В. Плеханова, 2009.
4. Михайлов А.И., Гиляревский Р.С., Черный А.И. Основы информатики. – М.: Наука, 1968. – 756 с.
5. Белоногов Г.Г., Котов Р.Г. Автоматизированные информационно-поисковые системы. – М.: Советское радио, 1968.
6. Белоногов Г.Г., Новоселов А.П. Автоматизация процессов накопления, поиска и обобщения информации. Библиотечка программиста. – М.: Наука, 1979.
7. Белоногов Г.Г., Кузнецов Б.А., Новоселов А.П. Автоматизированная обработка научно-технической информации. Лингвистические аспекты // Итоги науки и техники. Серия Информатика. Том 8 / под ред. В.И. Горьковой.– М.: ВИНИТИ, 1984.
8. Белоногов Г.Г. Теоретические проблемы информатики. Том 2. Семантические проблемы информатики / под общей ред. К.И. Курбакова.– М.: РЭА им. Г.В. Плеханова, 2008.
9. Черный А.И. Всероссийский Институт Научной и Технической Информации: 50 лет служения науке. – М.: ВИНИТИ, 2005.
10. Гиляревский Р.С. Информационный менеджмент: управление информацией, знаниями, технологией. – М.: Профессия, 2009.
11. Белоногов Г.Г., Хорошилов Ал-др А., Хорошилов Ал-сей А. Единицы языка и речи в системах автоматической обработки текстовой информации // Научно-техническая информация. Сер. 2. – 2005. – № 11. – С. 21-29.
12. Белоногов Г.Г., Хорошилов Ал-др А., Хорошилов Ал-сей А., Козачук М.В., Рыжова Е.Ю., Гуськова Л.Ю., Каким быть машинному переводу в XXI веке // Сб. “Перевод: традиции и современные технологии”. – М.: ВЦП, 2002.
13. Белоногов Г.Г., Гиляревский Р.С. Еще раз о гносеологическом статусе понятия “информация” // Научно-техническая информация. Сер. 2. – 2010. – № 2. – С.1-6.
14. Белоногов Г.Г., Гиляревский Р.С., Хорошилов А. А., Хорошилов-мл. А. А. Автоматическое распознавание смысловой близости документов // Научно-техническая информация. Сер. 2. – 2011. – № 7. – С. 15-22.
15. Белоногов Г.Г., Гиляревский Р.С., Хорошилов А. А. Проблемы автоматической смысловой обработки текстовой информации // Научно-техническая информация. Сер. 2. – 2012. – № 11. – С. 31-38.
16. Новая философская энциклопедия. Т. 1-4. – М.: Мысль, 2000.
17. Максименко С.Д. Общая психология. – Киев: Рефл-бук, Ваклер, 2000.
18. Соссюр Ф. де. Курс общей лингвистики. – М.: Прогресс,. 1977.
19. Спиркин А.Г. Философия (2-е издание). – М.: Гардарики, 2006.
20. Звегинцев В.А. Предложение и его отношение к языку и речи. – М.: Изд-во Московского университета, 1976.
21. Nagao M. A framework of a mechanical translation between Japanese and English by analogy principle, in Artificial and Human Intelligence / ed. A. Elithorn, R. Banerji. – North Holland, 1984. – P. 173-180.
22. Webb L. E. Advantages and Disadvantages of Translation Memory: a Cost/Benefit Analysis. – San Francisco State University, 1992.
23. Жуков Д.А. Мы переводчики. – М.: Знание, 1975.
24. Белоногов Г.Г., Хорошилов Ал-др А., Хорошилов Ал-сей А.. Автоматизация составления англо-русских двуязычных фразеологических словарей по массивам двуязычных текстов (билингв) // Научно-техническая информация. Сер. 2. – 2010. – № 5. – С. 1-8.

ПРИЛОЖЕНИЕ

Фрагмент словаря синонимических словосочетаний

Словарь составлен путем извлечения из англо-русского политематического словаря системы RETRANS синонимических последовательностей русских переводных эквивалентов английских фразеологических словосочетаний. Объем словаря – более 290 тыс. словарных статей.

1. моторизованные войска / моторизованные силы
2. центр боевого управления / центр управления боевыми действиями
3. секретный документ / конфиденциальный документ / документ, не подлежащий оглашению
4. авиационно-космическая медицинская подготовка / подготовка по авиационной и космической медицине
5. бензиновый грузовой автомобиль / грузовой автомобиль с бензиновым двигателем
6. полет вне установленного маршрута / внетрассовый полет
7. разведывательная аппаратура / разведывательное оборудование
8. тормозная двигательная установка / тормозной двигатель
9. отбор члена авиационного экипажа / отбор члена летного состава
10. наземная контрольно-испытательная аппаратура / наземная проверочная аппаратура
11. неорганизованная атака / беспорядочное наступление
12. государственные средства / ресурсы государства
13. атака мятежников / нападение мятежников
14. трудная задача / трудная проблема
15. злоумышленное использование / злоумышленное применение
16. ряд колебаний / серия колебаний
17. бактериальная инфекция / бактериальное заражение / бактериальные болезни
18. законопослушный человек / лицо, соблюдающее закон
19. излучение внеземного происхождения / внеземное излучение
20. я не совсем понимаю вашу мысль / я не совсем понимаю, что вы имеете в виду
21. перспективная ракета / ракета будущего поколения / ракета следующего поколения
22. потери во входном устройстве / потери на входе
23. только один раз / только однажды
25. внесудебная мера пресечения / внесудебное лишение свободы / внесудебный запрет
26. зимний период / зимнее время
27. робот, работающий в декартовой системе координат / робот, работающий в прямоугольной системе координат
28. на глаз / на ощупь
29. пеленгаторный приемник / пеленгационный приемник
30. солнечный коллектор с собственным энергообеспечением / солнечный коллектор с собственным источником питания
31. фильтр с очень мелкой сеткой / фильтр тонкой очистки
32. диаграмма рассеяния / индикатриса рассеяния
33. сохранять спокойствие / оставаться спокойным / не терять голову / сохранить хладнокровие / быть спокойным / не волноваться
34. изложить свои идеи в книге / сформулировать свои идеи в книге
35. неполная очистка / частичная очистка
36. калориметрическое испытание / калориметрическая проба
37. персональная аналоговая ЭВМ / персональный аналоговый компьютер
38. культура, выращиваемая на экспорт / экспортная культура
39. приводить в порядок дела / устраивать дела
40. создать правовую норму / создать закон
41. открывать дебаты / начинать дебаты
42. входит в состав / является частью / составляет часть
43. отсос пограничного слоя / отсасывание пограничного слоя
44. пропади оно пропадом / тьфу, пропасть
45. цитопатогенный эффект / цитопатогенное действие
46. годовой доход / ежегодный доход
47. волновод поверхностных акустических волн / волновод ПАВ
48. достичь зрелости / достичь полного развития
49. РЛС для обнаружения метеорных следов / радиолокационная станция для обнаружения метеорных следов
50. вот те на / вот так-так / вот так штука / вот это да / подумать только / ну и ну
51. испытание на термостойкость / испытание на теплоустойчивость
52. рулевая машина / силовой привод руля
53. передовой / находящийся на передовой линии
54. выгрузное устройство / разгрузочное устройство
55. она была так слаба, что еле шла / она была так слаба, что шла с трудом / она была так слаба, что едва могла двигаться
56. устройство для считывания штрихового кода сканер штрих-кода
57. станок для вставки петель / станок для установки петель
58. справедливо, что / справедливости ради следует отметить
59. латентная инфекция / дремлющая инфекция
60. малярная мастерская / малярный цех / окрасочный цех
61. формирование объемных изображений / трехмерная визуализация
62. невзирая на дождь / несмотря на дождь
63. понести тяжелый урон / сильно пострадать
64. научные исследования / научные разработки
65. реле перегрузки / реле максимального тока / реле защиты от перегрузок

66. с той разницей, что / с той лишь разницей, что / с той только разницей, что
67. как и ожидалось / как и следовало ожидать
68. ведущая ось / ось приводного колеса
69. в общем он прав / в целом он прав
70. дерево в возрасте плодоношения / плодоносящее дерево
71. очистка газов высокой температуры / очистка горячих газов
72. по обе стороны от / с каждой стороны
73. специальная аппаратура / специальное оборудование / спецоборудование
74. срок действия / срок годности
75. убивать время / коротать время
76. быть на рассмотрении / обсуждаться
77. низшая теплотворность / низшая теплотворная способность / низшая теплота сгорания
78. в духе согласия / в конструктивном духе
79. эта шляпа вам очень идет / эта шляпа вам очень к лицу
80. просить о разрешении / просить разрешения
81. проезд закрыт! / прохода нет!
82. давать показания / свидетельствовать / давать свидетельские показания
83. я сделаю это, если смогу / я постараюсь это устроить
84. быть увешанным флагами / быть украшенным флагами
85. система распыления воды / водораспылительная система
86. генеральная политика / общая политика
87. особые условия страхования / специальные условия страхования
88. сердечно-сосудистая реактивность / реактивность сердечно-сосудистой системы
89. вид разрушения / тип разрушения
90. я предлагаю его исключить / я вношу предложение его исключить
91. мне надоели ваши " если " и " но " / мне надоели ваши сомнения и возражения
92. передача тепла радиацией / теплообмен излучением / радиационный теплообмен
93. осаждение тонкого слоя / нанесение пленки
94. полиномы Чебышева / многочлены Чебышева
95. неорганизованная атака / беспорядочное наступление
96. биологические влияния / биологическое воздействие
97. строительная площадка / стройплощадка
98. кровяное пятно / пятно крови

Материал поступил в редакцию 11.03.13.

Сведения об авторах

БЕЛОНОГОВ Герольд Георгиевич – доктор технических наук, профессор, главный научный сотрудник ЗАО RETRANS Technologies, Москва
e-mail: belonogov@mail.ru

ГИЛЯРЕВСКИЙ Руджеро Сергеевич – доктор филологических наук, профессор, зав. Отделением ВИНИТИ РАН, Москва
e-mail: giliarevski@viniti.ru

СЕЛЕТКОВ Сергей Николаевич – доктор технических наук**,** профессор Московского экономико-статистического института

ХОРОШИЛОВ Александр Алексеевич – доктор технических наук, ведущий научный сотрудник Института проблем информатики, Москва
e-mail: khorochilov@mail.ru

УДК 316.776.23:002.1

Е. И. Полтавская

Информация субъективная, социально-опредмеченная и документ*

Сравнивается субъективная и социально-опредмеченная информация. Документ рассматривается как материальный предмет, который способствует уменьшению неопределённости при принятии управленческого решения в рамках конкретной коммуникации.

Ключевые слова: *субъективная информация, социально-опредмеченная информация, документ, идеальное, материальное*

Любое знание субъектно и субъективно (есть познающий субъект, индивидуально интерпретирующий поступающие сигналы), объектно (наличествует объект исследования, который «помещается» вне познающего субъекта; даже познавая самого себя, свой внутренний мир, субъект изучает его *со стороны*, как бы вне своего сознания) и объективно (реально, независимо от сознания субъекта). Кроме того, существует общее социальное знание (в рамках которого с рождения оказывается человек), которое Э. В. Ильенков считал истинно идеальным и объективным (проверенным деятельностью) [1], но которое по большей части является опредмеченным, поскольку из субъективного идеального в мозге индивида перекодировано на другой материальный носитель вне субъекта. Это общее социальное знание, конечно, имеет не только объективные, но и субъективные черты: заблуждения бывают и коллективными. Поэтому, если говорить об источниках информации, индивид субъективно осмысливает явления природной и социально-культурной сфер и, наконец, собственный субъективный мир.

Проблемы индивидуального сознания и личностного начала в общественном сознании, соотношение понятий «идеальное» и «информация» много лет разрабатывает Д. И. Дубровский. Информация понимается им как «содержание отражения на уровне самоорганизующихся систем» и эмерджентный результат – следствие усложнения системы, когда свойства составляющих её элементов значительно отличаются от общей системной функции. Опираясь на достижения естественно-гуманитарного знания, Дубровский смог связать содержательно-ценностное описание информации с описанием её кодовой зависимости от материального носителя и тем самым объединить достижения естественных и гуманитарных дисциплин, подтвердив общенаучный уровень категории информации. Причём, в отличие от Ильенкова, Дубровский считает идеальной только субъективную реальность [2, § 4.2].

Из четырёх основных форм существования информации, выделяемых Дубровским (*допсихическая*; *психическая*[1]; *анимально-опредмеченная* и *социально-опредмеченная)*, в рамках решаемой проблемы интересны два вида: психическая и социально-опредмеченная. Психическая форма существования информации включает субъективную реальность человека и животных с учетом их качественного различия. Учитывая доводы Дубровского, что понятие «психическое отражение» в основном включает «отражательные акты, которые совершаются в форме субъективных образов и состояний», что термин «психическое отражение» относится к психологическим, а не философским терминам [там же, § 2.1] и что информация – это всякое явление субъективной реальности, будем называть информацию, создаваемую человеком, *субъективной*. Социально-опредмеченная информация – это результаты деятельности человека. Можно сказать, что *социально-опредмеченная информация* – это воплощённый материально замысел человека-творца, который распредмечивается другими людьми.

Осмысление нельзя считать чем-то автоматическим, безусловным. На него влияют множество факторов, из-за которых человек, читающий в разное время, например, одну и ту же книгу, может прийти к разным выводам, испытать разные чувства. Сознание вообще имеет свойство интенциональности, т. е. нацеленности на какой-то предмет, и поэтому субъективная реальность всегда имеет какое-то содержание.

Что можно узнать из литературного произведения, например, «Война и мир»? Один сможет представить себе жизнь людей в определённое время в определённой культуре и в определённых обстоятельствах: о чём думали, мечтали и что совершали типичные или выдающиеся люди, каковы их горести и радости, устремления и чаяния; узнать об их вере и заблуждениях, добре и зле. Другой – почувствовать и осознать дух того времени, свободу сознания как «подчинение его необходимому ходу вещей». А кому-то интересно выяснить влияние личности автора на текст. Иными словами, материалистической философией признается факт обусловленности человеческого сознания оп-

* См. Научно-техническая информация. Сер. 2. – 2013. – № 5. – С. 1–6.

[1] Допсихическая и психическая информации составляют, полагаю, когнитивную информацию.

ределенным кругом бытия, но допускается возможность развития духа [3].

В терминах информационного подхода всё вместе можно назвать информацией, преломлённой в мировоззрении автора, или *субъективной информацией об объективном и субъективном социальном мире*.

Существует представление, что феномен информации есть результат *дуализма* реальности, взаимодействия между собой материальных и идеальных компонентов (позиция К. К. Колина, В. Л. Обухова, А. В. Соколова).

Информация, по Колину, относится к идеальной реальности. Причём идеальная реальность, по его мнению, существует объективно, «независимо от деятельности сознания», наравне с физической реальностью и возникает при взаимодействии объектов физической реальности. При этом свойства одних объектов (или процессов) отражаются в структуре других объектов (или процессов), создавая феномен информации [4, с. 71, 75]. Таким образом, Колин руководствуется теорией отражения и положительно решает вопрос о тождестве мышления и бытия (т. е. предполагает адекватное отражение действительности с помощью мышления).

Марксизм-ленинизм, как известно, также признавал идеальное (сознание, бессознательное, мышление) – но как внутренний мир человека, субъективную реальность, а всё объективно существующее относил к проявлению материальной реальности: «в мире нет ничего, что не было бы материей, её свойством или каким-либо продуктом её развития» [5, с.31]. Трактовка идеального как субъективной реальности подчёркивает *реальность* мыслей, чувств и их принадлежность *субъекту*, а не внешнему миру (но субъективная реальность Другого для меня является объективной). По современным материалистическим представлениям мышление осуществляется с помощью внутренних ментальных репрезентаций информации в когнитивной системе. «Эта когнитивная репрезентация мысли не является синтаксической, подобно символьному языку, она также не локализована в отдельных нейронных узлах или нейронах, а *распределена* в системе (курсив мой. – *Е. П.*)» [6, с. 251.].

Если логические операции, выполняемые компьютерной программой, не эквивалентны устройству компьютера («железу») и к нему не редуцируемы, то естественно предположить, что явления, происходящие в субъективной реальности, не следует сводить к строению мозга (они, однако, зависят от его сложности). В таком случае фразу Колина «Идеальная реальность объективно существует независимо от деятельности сознания», возможно, следует понимать не как свидетельство появления ещё одной идеальной реальности помимо сознания, а так, что идеальное реально существует, проявляясь как свойство высокоразвитой материи – мозга субъекта. Идеальное (например, впечатление от вещи) может образовываться вне зависимости от сознания индивида, поскольку может возникнуть *неосознанно*, в виде перцептивных репрезентаций, схем [6, с. 253, 260], а потом проявиться как интуитивное понимание (восприятие сплетено с мыслительным процессом). Кроме того, мысленный образ не зависит от воспринимающего сознания и в том смысле, что он фиксирует реальные отношения, ту форму деятельности, в которой была создана вещь, и которая действительно независима от субъекта. Например, схема логической формы понятия не является следствием текущего состояния индивида (думаю, в этом смысле Ильенков и говорит об идеальном как *о законе*, управляющем сознанием и волей человека).

Однако само идеальное формируется в субъективном мире, поэтому объективное влияние на формирование идеального оказывает не только реальная предметная социально-культурная деятельность человека (проявляемая как коллективное сознание), но и активность его сознания (например, при мышлении, целенаправленном восприятии). Всякое явление сознания, согласно Дубровскому, есть определенная информация. Информация, доказывает он, всегда об объекте, всегда чья-то (а именно субъекта) и всегда связана с материальным носителем – воплощена в конкретном коде: идеальное как отдельная реальность материалистами отрицается. Следовательно, согласно материалистическим воззрениям, хотя информация в субъективном мире человека идеальна, следует помнить о её неразрывной связи с *высокоорганизованным материальным носителем* информации.

Школа Обухова утверждает дуализм реальности, который проявляется в единой субстанции материи и духа (монодуализм) – как их диалектический синтез, «с двумя противоположными ликами, несводимыми один к другому» [Цит. по: 7, с. 158], т. е., *единодвойственность*. Мир представляется эволюционирующим целым, «где нет тождества и нет разрыва между материальными и духовными составляющими, но есть различного рода взаимодействие между ними на различных уровнях бытия» [8, с. 457]. Эта философская концепция, лишенная недостатков любого материализма, которому трудно обосновать возникновение идеального в материальном, также имеет изъян. Главный недостаток монодуализма видится в том, в чём постулируется его достоинство. В соответствии с этой гипотезой, «материя и дух *всегда* интегрированы, *всегда* дополняют друг друга (курсив мой. – *Е. П.*)», они нераздельны, подчёркивает Соколов, как две стороны одного листа [7, с. 157]. Однако именно это и противоречит реальности: *идеальное порождается лишь в высокоорганизованной материи* – в самоорганизующейся системе (в человеке – как его субъективный мир; у некоторых высших животных; возможно, в новейших искусственных системах). Именно человек *видит* наличие информации везде и всюду, употребляя для быстроты коммуникации метафоры, *приписывая* информации *субстанциональность*, в то же время зная, что она (информация) не обнаружена (и, наверное, не может быть обнаружена). Вопрос, как «сочетаются» материя и дух в одной субстанции, не исчезает и в монодуализме.

Исследование концептуальной природы информации является одной из актуальных фундаментальных проблем науки. Учению об информации посвящены труды философов и учёных разных отраслей знания

[9]. Среди многочисленных попыток разобраться, что такое информация, – работы отечественных специалистов в области документо-коммуникационных наук разной степени углубления в проблему (от философского осмысления сущности информации до применения информационного подхода в библиотечном деле): Ю. Н. Столярова [10], А. В. Соколова [7], Т. Ф. Берестовой [11], М. Я. Дворкиной [12]. Остановимся на некоторых моментах недавнего труда библиографоведа и библиотековеда, информатика и философа А. В. Соколова, который глобально и оригинально подошёл к осмыслению информации.

Этот ученый, приверженец «реалистической философии» В. Л. Обухова, дал определение разных видов информации. Важнейшим свойством информации, по Соколову, является амбивалентность – двойственность, выражающаяся в её материально-идеальной природе. Поскольку наши исходные точки зрения на бытие значительно расходятся, это, естественно, сказывается и на представлении об информации. Поэтому по многим вопросам дискуссия бессмысленна, так как упирается в наше различное понимание, что такое «монодуальность» мира. Например, как и Колин, Соколов считает информацию идеальной: «информацию нужно считать идеальным образованием (даже смыслом в чистом виде)» [7, с. 252]. В таком случае, возникает вопрос о пропагандируемом им монодуализме: если информация идеальна, то где же материя, с которой она должна быть *неразделима* как стороны одного листа?

В связи с целями настоящей статьи интересно обсудить информацию в самоорганизующихся системах, поэтому обратим внимание на *сущностное* определение семантической информации, данное Соколовым: это «*амбивалентный феномен, выражающий духовные смыслы в коммуникабельной знаковой форме*» [там же, с. 253].

Семантическую информацию автор считает «*выражением* идеального» [там же], следовательно, он понимает её как опредмеченное. Чтобы субъективное идеальное автора стало доступным другим людям во времени и пространстве, идеальное надо *опредметить*, сделать материальным: воплотить замысел в книгу, картину, в ноты или звуки, в научную статью или устный доклад, техническое устройство. Без материального сделать своё идеальное доступным для других невозможно. В этом смысле вся культура есть *выражение идеального* – имеет дуальное происхождение. В том и состоит задача творческого человека – найти такую материальную форму, чтобы Другой, увидев/услышав/прочитав произведение, осознал то содержание, которое в него вложено автором. Гениальному художнику удаётся придать материальному предмету такую форму, которая создаёт множество разнообразных образов, ассоциаций у людей в разные эпохи. Чем богаче личность воспринимающего субъекта, тем больше его «веер» интерпретаций, причём зачастую интерпретатором не только воссоздаётся творческий замысел, но и добавляется новый смысл, подчас не осознававшийся изначально самим автором (потому-то так интересно слушать одно и то же музыкальное произведение в разном исполнении или смотреть разные постановки пьесы, содержание которой известно).

Если принять точку зрения Дубровского, становится понятно: идеальное не может содержаться в предметах культуры, т. к. они объектны, находятся вне субъективного мира человека, где только и может возникнуть идеальное. Книга, какое бы духовное содержание она ни несла, материальна: это только блок бумаги, испещрённый знаками. Но у подготовленного человека (знающего кодирующую систему знаков, в данном случае язык) сочетание этих знаков вызывает разнообразные мысли и чувства, может побудить к действию.

«Семантическая информация» как выражение идеального есть синоним *опредмеченного*. То есть это вовсе не амбивалентное, не идеальное, а настоящее материальное: книги, картины, музыкальные произведения и т. д. Кстати, если бы те же книги имели амбивалентный, а не материальный характер, то было бы невозможно выражение «рукописи не горят» (в смысле неуничтожимости их смысла), и с каждым уничтоженным экземпляром книги исчезало бы идеальное.

Идеальное содержание, вложенное автором, может быть с разной степенью адекватности *представлено* только другим *субъектом*, для чего нужно также обладать определённым запасом знаний, фантазией, целями, проявлять активность в познании, обладать волей и т. д. Поэтому, действительно, опредмеченная информация есть целенаправленное *выражение* духовного содержания с помощью материального. Причём творческий замысел не полностью совпадает с распредмеченным – воспринятым другим субъектом.

Термин «семантическая информация», на мой взгляд, только запутывает дело, не добавляя ничего нового по сравнению с «опредмеченным». Ведь если информация (как идеальное) всегда интенциональна (предметна, имеет содержание), то она должна быть и семантической. Это выражено в определении информации Ю. Н. Столярова: «информация – это семантическое преобразование отражения реальности субъектом живой природы» [10, с. 50]. Однако про опредмеченную («семантическую») информацию лучше говорить, что она имеет содержание, поскольку признак интенциональности относится только к субъективной реальности человека. *Интенциональность*, как верно отметил Соколов, означает не просто вторичность, а обусловленную внутренним, субъективным, смысловую направленность, устремлённость на объект (от *intention* – замысел, намерение) [7, с. 249]; «полагание предмета в мысли» – свойство, присущее лишь человеческому сознанию [13]. Именно поэтому *вторичность* и *интенциональность* не являются синонимами, их не стоит заменять друг другом. Понятие «информация», как оно мыслится сторонниками разных концепций, является собирательным и не является общим по отношению к понятию «семантической информации» (овеществлённому, материальному) – эти понятия не могут находиться в соотношении род–вид.

Трактовка информации как феномена и феномена как сущности [7, с. 250] вызывает сомнения из-за неоднозначности термина «феномен», значение которого различно у разных авторов – от Платона (у него

оно менялось на диаметрально противоположное: и «кажимость», и «положительная явленность» бытия) до Хайдеггера («феномен не обозначает ничего “содержательного”») [14]. Если опираться на исходное толкование феномена в «Новой философской энциклопедии», то феномен – *не сущность*, а выдающееся *явление.* Феномен – это то, что явно, т. е. видно, заметно с помощью органов чувств, что *ощущается*. Таким образом, феномен – явен для наблюдателя, а сущность всегда скрыта, её выявляют. Сущностное повторяется во всех предметах определённого рода, оно устойчиво. Феномен – как всё яркое, исключительное, особенное – редок. Коль скоро сущность – это «та сторона индивидуального предмета, которая определяет все другие его стороны» [15, с. 946], то феномен – «явление, предмет, данный в чувственном созерцании» [14].

Подставим толкование феномена из «Новой философской энциклопедии» в формулировку «информация есть феномен» и получим: «информация есть явление, предмет, данный в чувственном созерцании». Однако, если информация есть «предмет, данный в чувственном созерцании», то это противоречит тому, что отсутствуют какие-либо прямые или косвенные опытные доказательства существования информации: информация как раз и не обнаружена в чувственном созерцании как некий предмет (о чём пишет и сам Соколов) [7, с. 14; см. также: 9, с. 6].

Сущность можно познать в чувственных восприятиях с помощью логических умозаключений, абстрактного мышления, в том числе с помощью схем и других интеллектуальных построений. Сущность, как и «феномен», понимается в феноменологии как идеальное образование, присущее субъективной реальности. Поэтому толкование «феномена» Э. Гуссерлем отличается от современного общепринятого и слово приобретает другое значение. Гуссерль пишет: «Феномен – акт “чистого сознания”, имеющий идеальную природу» [Цит. по: 7, с. 250]. Феномен, по мнению автора, это внутреннее впечатление от внешней характеристики предмета, его осознание, т. е., феномен – это постижение предмета в виде *образа сознания*. Отсюда выражение «информация есть феномен (чего-то)» по Гуссерлю можно трактовать как «информация (о чём-то) есть интеллектуальный образ (чего-то)». Но идея о том, что интеллектуальный образ идеален – несовместима с заявленной монодуальностью. Поэтому использование термина «феномен», даже с добавлением «амбивалентный», в определении информации, в том числе семантической, не проясняет их сущности. Не привносит ясности и комментарий, где указывается, что «“амбивалентный феномен” *не следует понимать как “идеально-материальный объект“*, потому что феномен – акт “чистого сознания” и никаких материальных включений не содержит (курсив мой. – *Е. П.*)», несмотря на то, что в предыдущем абзаце написано обратное[2] [7, с. 274]. Из сказанного следует, что Соколов не последователен в приверженности к монодуализму.

Интерпретация формы и содержания интеллектуальных конструктов идеальной природы, «выстраиваемых» субъектом сознательно в процессе его жизнедеятельности в зависимости от культуры, которой он обладает и к которой принадлежит, создаёт у него представление о наличии информации в природных, социальных, био-, и социотехнических системах и впечатление дуальности реального мира. Сигналы объективного мира вызывают в живом организме «субъективные следы, остающиеся в памяти» [16, с. 80], – образы (феномены, по Гуссерлю). Образ можно схематизировать, осмысливая явление, процесс или вещь, представленную «в замедленной оптике» как процесс создания предмета, и тогда постичь его сущность.

Наиболее актуальные образы, позволяющие на данном этапе развития общества объяснить некоторые процессы восприятия и понимания, память, представить информационное единство мира и др., опредмечиваются и становятся коллективным достоянием, влияющим на индивидов.

На *реальном*, но *идеальном* существовании интеллектуальных конструктов – схем в мозге субъекта, которые затем закрепляются в понятии, а далее в дефиниции (и тем самым кодируются на другом материальном носителе), чтобы полученное знание не пропало, не ушло вместе с индивидом, а могло быть усвоено и применено другими, основан метод схематизации понятий социальных общностей, развиваемый автором данной статьи [17].

Изменения во внешнем (или внутреннем) мире, которые воспринимает и интерпретирует человек, принято называть *сигналом* и расценивать его как материальный «носитель информации». Слово «сигнал» (от лат. *signum* – знак) в основных европейских языках толкуется как «знак, обозначение». Но понятие «сигнал» шире, чем понятие «знак», поскольку сигнал включает и знак (условный заместитель некоего сообщения), и неспециальную форму передачи информации. Так, увидев за окном гнущиеся от ветра деревья, мы можем судить о погоде. Сигналы об изменениях реальности безадресны и доступны одновременно многим. Только каждый субъект воспринимает их по-своему, т. е. субъективно. Иногда высказывается мнение, что бессмысленный набор звуков, действий не является семантической информацией [7, с. 253]. Однако этот бессмысленный набор сигналов может свидетельствовать о состоянии человека/машины, отправляющего сообщение: т. е. то, что кажется информационным шумом, может сопровождаться осмысленной информацией для принимающей стороны [16, с. 81].

Знак – это *специально выбранный* сигнал общения, который заменяет обозначаемую величину, он произволен, возникает как результат договорённости [18, с. 20].

[2] «Семантическая информация – *амбивалентный* [идеально-материальный] *феномен* [акт сознания, доступный вычленению и относительно самостоятельному изучению], *выражающий духовный смысл* [субъект-объектное отношение, включающее понимание значения объекта, оценку его ценности для субъекта и мотивацию реакций субъекта на данный объект] *в форме коммуникабельных знаков* [средство выражения смысла, рассчитанное на восприятие реципиентом]».

Идеальный замысел фиксируется знаками, поэтому признак «*знаковая форма* выражения смыслов» характеризует опредмеченное (в том числе документ). Но субъективная информация может не быть знаковой (мышление без слов): при понимании, восприятии, отражении, образном мышлении, интуиции. Заметим, что в перечне основных свойств информации, с которыми согласны многие специалисты, знаковая форма не указана [16, с. 80–81], видимо потому, что субъективная информация может быть незнаковой.

По Дубровскому, «внесловесная мысль существует, она объективирована в мозговых нейродинамических системах – кодах определенного типа, отличных от кодов внутренней речи; она представляет собой специфическую разновидность и неотъемлемый компонент субъективной реальности» [2, § 2.3]. Незнаковая информация выражается в понятии, чувстве, визуальных представлениях, сценариях, в виде абстракций [6, с. 253–254; 19, с. 20]. Существует гипотеза, что информация может восприниматься не только в процессе разложения на составляющие свойства, но целостно – из-за того, что «глаз как орган мозга способен принимать не только классические состояния входящего света, но и непосредственно квантовые состояния фотонов» [20, с. 95]. Можно предположить, что возможность восприятия информации целостно – признак не только высокой скорости образования интеллектуальных образов, но именно незнаковой формы информации. О том, что информация может быть выражена незнаково, свидетельствуют, например, эксперименты К. Малевича, когда было выявлено объективное влияние разного цвета на ощущения человеком формы предмета и его габаритов, размеров пространства, влияние на настроение, здоровье. Владея этим знанием, художник может передавать своё восприятие жизни другим и влиять на их ощущения, т. е. специально оперировать сигналом как знаком [21, с. 66–77].

И. В. Мелик-Гайказян, разделяя атрибутивный подход и видя информацию везде, представляет ее в виде *процесса.* Она основывает свою гипотезу на осмыслении взглядов А. Н. Уайтхеда (изложенных им в книге «Процесс и реальность») и достижениях в нейросемантике (направление психологии) [22, с. 100].

В нейросемантике разработано интегрирующее определение информационного ресурса: «информации» и «знания». Суть его изложил В. И. Бодякин [там же, с. 104–108]. Согласно этой гипотезе, физические системы для своего выживания вырабатывают информационно-управляющие системы (ИУС). Между окружающим миром и ИУС происходит *процесс* редукции описания внешней среды (объекты также представляются как процессы) с помощью разной последовательности знаков. В результате процесса редукции *внутри* ИУС создаётся информационная модель (образ) предметной области: возникает информация. Таким образом, информация порождается ИУС в ответ на изменения во внешней среде; модель информационного ресурса позволяет представить *информацию* как «простую взаимно-однозначную функцию (“один образ – один процесс”)», а формирование *знания* как «множественный процесс (“один образ – класс процессов”)» [22, с. 108].

Удовлетворимся принятыми положениями: информации нет без человека и высших животных. Для информации, определяемой как управляющий сигнал (субъективная информация), и социально-опредмеченной информации характерно не только разное существование, но и разные свойства. Если первая «уменьшает неопределённость системы», суживает варианты понимания, то вторая обладает другим свойством: облегчает социализацию, а при необходимости выбора создаёт больше возможностей, даёт свободу самовыражения. Материалистами принимается, что *субъективная информация* всегда связана с материальным носителем – мозгом – и *называется идеальной,* поскольку в мозге нет «физического отпечатка отражения. Есть лишь его внутренние репрезентации» [23, с. 60], которые индивидом воспринимаются каким-то образом целостно.

В соответствии со сказанным, приведём рабочую формулировку своего осмысления субъективной информации: *Субъективная информация – это процесс создания индивидом идеальных ценностно-смысловых проекций картин мира (объекта) при действительном или мысленном контакте с ним и при общении с другими субъектами*.

Поясним. Субъективная информация представляется процессом потому, что формирование комплексов достраивающихся и перестраивающихся конструктов иногда значительно растянуто во времени (на что указывает присущая человеку рефлексия – осмысление и оценка своих собственных действий, а также медленное понимание, которое обозначается выражением «дошло как до жирафа»): «содержание» сознания всё время преобразуется, мышление находится в движении.

В соответствии с синтаксической, прагматической и семантической характеристиками информации, по Дубровскому, проекция объекта – это интеллектуальные конструкты: образы, понятия, схемы, или *форма* поступившей информации, выражающая её ценность и смысл для субъекта. В картину мира (объект) входят конструкты не только окружающей среды, но и сам субъект, рассматриваемый как бы «со стороны».

Интерсубъективность субъективной информации проявляется в том, что субъект живёт и развивается в обществе, зависит от общения с другими людьми, с другим *Я*. Поэтому на субъективную информацию оказывает большое влияние социальное общение, она корректируется под влиянием субъективности других субъектов и обобщается.

Если субъективная информация перекодирована, т. е. выражена знаками на любом другом носителе кроме мозга – это *опредмеченная информация*. Устную речь, как и музыкальное произведение, видимо, тоже следует отнести к социально-опредмеченной информации, поскольку их можно представить как созданный предмет-процесс, при котором идеальное содержание выражается звуковыми волнами (знаками).

В документо-коммуникационных науках имеют дело не с информацией – *субъективным идеальным*, а только с *социально-опредмеченной информацией* (со-

циально-опредмеченным знанием), т. е., с тем, что в общенаучном (нематематическом) плане именовалось информацией и трактовалось как получение «новых сведений об объекте» [24, с. 142]. Термин «социально-опредмеченная информация», думаю, поможет различить, о какой информации речь – информации как идеальном, присущем лишь субъекту, или информации, воплощённой человеком в материальном. Сравним признаки субъективной и социально-опредмеченной информации в таблице.

Феномен социально-опредмеченной информации необходим, чтобы знание могло быть социализировано, превращено в общественное достояние. Так выстраивается причинно-следственная цепочка: внешний или внутренний *сигнал*, воспринятый субъектом – возникновение *субъективной информации* (идеальных образов, схем) – *социально-опредмеченная информация* (замысел, воплощенный в материальном продукте) – внешний *сигнал* для Другого – «усвоение» информации Другим и переработка её в виде собственного *знания* (образов, схем). Социально-опредмеченная информация есть лишь намёк на знание, по которому воспринимающий индивид должен воссоздать само знание *в себе*, поскольку знание личностно (о соотношении информации и знания см. [10, с. 64–66]).

Социально-опредмеченная и субъективная информации – это результаты двух взаимосвязанных процессов: опредмечивания и распредмечивания, которые можно рассматривать как единый цикличный процесс. Процесс опредмечивания-распредмечивания сопровождает процесс образования социального института – коммуникацию минимум двух субъектов: созидателя и потребителя. Действительно, любой социокультурный предмет есть результат воплощения идеального авторского замысла в материальном (опредмечивание). Созданный предмет предлагается потребителю, при этом осуществляется коммуникация *созидатель–потребитель*. При восприятии, осмыслении его субъектом-потребителем происходит обратный процесс – распредмечивание, воссоздание модели авторского замысла в той мере, в какой на это способен потребитель. Процесс опредмечивания и распредмечивания можно представить как взаимосвязанный «обмен социально-опредмеченно-распредмеченной информацией» – *преобразование субъективного ценностно-смыслового содержания (из кода мозговой ткани) и воплощение его субъектом-созидателем на любом материальном носителе с последующим расшифровыванием субъектом-потребителем.* Социально-опредмеченно-распредмеченная информация, сама являясь процессом (превращения идеального в материальное и обратно), сопровождает любой социальный процесс или коммуникацию между субъектами, выполняющими разные взаимодополняющие и взаимозависимые функции созидателя и потребителя.

Таблица

Субъективная информация	**Социально-опредмеченная информация**
Идеальна	Материальна
Возникает в высокоорганизованной материи – мозге человека или высшего животного; предположительно может быть создана в искусственных системах	Создаётся человеком. Референты – любые материальные предметы (вещь, процесс)
Интенциональна и субъективна (зависит от уровня культуры субъекта)	Вторична и объективна (реальна)
Выражается: знаком (внутренняя речь) и незнаково – образ, понятие, чувство, абстракция	Выражается: материальным сигналом, знаком, следом
Индивидуальна и в то же время интерсубъективна: зависит от физического и социального опыта, корректируется под влиянием других субъективных мнений	Социальна, коммуникативна
Уточняет знание, суживает варианты понимания	Облегчает социализацию, даёт свободу самовыражения
Не отчуждается от субъекта	Отчуждается от субъекта в виде устной речи или созданного предмета (вещи, процесса)
Субъектна (внутри живого субъекта)	Объектна (находится *вне* той системы, которая его воспринимает и анализирует)
Характеризует сознание индивида	Характеризует коллективное сознание
Формирует личностное знание индивида, исчезающее вместе с его смертью	Создаёт социально-культурное знание, передаваемое в процессе социализации и обучения последующим поколениям
Возникает в результате распредмечивания	Возникает в результате опредмечивания

Конечно, представления о социально-опредмеченно-распредмеченной информации схематичны и нужны лишь для различения сторон единого информационного процесса. Действительность всегда сложнее любой схемы. Так, на информацию, которую формирует субъект, неминуемо влияет и его генетика, и те культурные ценности, которые он смог почерпнуть, живя в обществе. И. П. Меркулов справедливо отмечает, что «“субъективность” человека – это результат интеграции *генетической* и *приобретённой культурной информации*» [23, с. 62]. Причём «генетическую информацию» можно понимать как ещё одно – природное – *выражение* информации (рост и размножение живой материи – это преобразование одних материальных объектов в другие), которое представляет собой характерный признак *строения живой материи*, способствующий ее лучшему воспроизводству и выживаемости в окружающей среде. Культура выполняет в обществе роль, аналогичную генетическим кодам в живых организмах, обеспечивая воспроизводство общества как целостной системы.

Предложение образовать в Номенклатуре научных специальностей ВАК новый раздел *Информационные науки*, куда планируется включить и документо-коммуникационные науки, ведёт к необходимости уточнить и соотношение понятий «информация» и «документ».

Признаки семантической информации, сформулированные Соколовым, – *смысловое (духовное) содержание*, а также возможность *использования в социальной коммуникации* и *вторичность* (если её рассматривать как обусловленность содержания оригиналом – замыслом субъекта) – присущи любой опредмеченной форме идеального [7, с. 253]. Документ, казалось бы, должен относиться к социально-опредмеченной информации. Действительно, как правило, документ создаётся субъектами, живущими в обществе, с определёнными коммуникационными целями, поэтому все признаки социально-опредмеченной информации должны иметься и у него как признаки рода. С помощью документа приобретаются некие сведения, которые помогают подтвердить события, получить недостающее знание, укрепиться во мнении, чтобы принять управленческое решение.

Как отметил Столяров, «не всякая информация приобретает статус документа, но всякий документ содержит информацию» [25, с. 1]. Так, субъективная информация, не перекодированная из мозгового кода в другой материальный код, документом не является. В то же время устное свидетельство есть результат такой перекодировки и может считаться документом в некоторых коммуникациях (например, «честное слово» уважаемого человека, сказанное им прилюдно).

Однако не все документы – результат социально-опредмеченной информации: упомянем о «документе-животном», которое не является продуктом культуры, и, значит, не может быть отнесено к социально-опредмеченной информации. Столяров прав: носителем информации может быть человек, животное, камень, которые могут в каких-то ситуациях являться и *документом* [там же, с. 1–2], причём – носителем информации для *внешней* по отношению к ним системе (субъекту), а не сами для себя. Документ нужен в качестве сведений, уменьшающих неопределённость для принятия какого-либо решения. Сравним с одним из распространённых определений информации: *«информация – это некое идеализированное сообщение, уменьшающее или полностью исключающее неопределённость в выборе одной из нескольких возможных альтернатив»* [26, с. 3–4]. Получается, что документ в принципе отличается от информации (идеального) только своей материальной природой.

Как уже было сказано, информацию субъект может *найти* в любом природном и социальном объекте при его распредмечивании, ведь одним своим существованием этот объект может свидетельствовать о чём-то важном, необходимом. Значит, не только человека или животное, но вообще любой материальный предмет можно считать *документом* в том случае, если его бытование будет расценено как уменьшающее неопределённость сведений в конкретной коммуникации. Итак, под документом можно понимать *материальный предмет (вещь, явление, процесс), который способствует уменьшению неопределённости при принятии управленческого решения в рамках конкретной коммуникации*. Таким образом, документом в конкретных коммуникациях социального общества может быть материальный предмет *любой природы* и его можно представить как фрагмент возобновляющегося процесса *социально-опредмеченно-распредмеченной (и вновь опредмеченной) информации.*

СПИСОК ЛИТЕРАТУРЫ

1. Ильенков Э. В. Проблема идеального // Эвальд Васильевич Ильенков. – М.: РОССПЭН, 2008. – С. 153 – 214.
2. Дубровский Д. И. Проблема идеального. Субъективная реальность. – 2-е изд. – М.: Канон+, 2002. – 368 с. – URL: http://www.globalistika.ru/dubrovsky/nauchnye_texty/probl_ideal.htm (дата обращения 20.01.2013).
3. Лифшиц М. А. Диалог с Эвальдом Ильенковым (Проблема идеального). – М.: Прогресс-Традиция, 2003. – 368 с.
4. Колин К. К. Философские проблемы информатики. – М.: БИНОМ. Лаборатория знаний, 2010. – 264 с.
5. Мелюхин С. Т. Избранные труды: Наследие и современность. В 3 т. – Т. 2. Философская онтология. – М.: Издатель Савин С. А., 2010. – 455 с.
6. Меркулов И. П. Мышление как информационный процесс // Эволюция. Мышление. Сознание. – М. : Канон+, 2004. – С. 228 – 260.
7. Соколов А. В. Философия информации. – СПб.: СПбГУКИ, 2010. – 363 с.
8. Миронов В. В. Бытие как центральная категория онтологии. // Философия. – М.: Норма, 2009. – С. 435 – 459.

9. Алошти Х. Р. Философский взгляд на информацию и информационную технологию // Научно-техническая информация. Сер. 2. – 2012. – № 4. – С. 1 – 12.
10. Столяров Ю. Н. Сущность информации. – М.: ГПНТБ России, 2000. – 120 с.
11. Берестова Т. Ф. Законы формирования структуры информационного пространства и функции информации // Библиография. – 2009. – С. 32 – 47.
12. Дворкина М. Я. Информационное обслуживание: социокультурный подход. – М.: Профиздат, 2003. – 112 с.
13. Мотрошилова Н. В. Интенциональность // Новая философская энциклопедия. В 4 т. – М.: Мысль, 2010. – Т. 2. – С. 133 – 134.
14. Михайлов И. А. Феномен // Новая философская энциклопедия. В 4 т. – М.: Мысль, 2010. – Т. 4. – С. 174 – 175.
15. Левин Г. Д. Сущность // Энциклопедия эпистемологии и философии науки. – М.: Канон+, РООИ «Реабилитация», 2009. – С. 946 – 947.
16. Романенко В. Н., Никитина Г. В. Многозначность понятия информации // Философия науки. – 2010. – № 4 (47). – С. 75 – 99.
17. Полтавская Е. И. Проблема «социальный институт vs организация» и её решение с помощью схем понятий // Вестник Нижегородского университета им. Н. И. Лобачевского. Сер. Социальные науки. – 2012. – № 3 (27). – С. 132 – 137.
18. Спирова Э. М. Философско-антропологическое содержание символа. – М.: «Канон+»; РООИ «Реабилитация», 2012. – 336 с.
19. Григорьев В. А. Знак: основные определения // Научно-техническая информация. Сер. 2. – 2004. – № 11. – С. 12 – 22.
20. Петренко В. Ф. Базовые метафоры как геном (зародыш) будущей теории (на материале психологической науки) // Вопросы философии. – 2012. – № 4. – С. 87 – 98.
21. Малевич К. Форма, цвет и ощущение // К. Малевич. Чёрный квадрат. – СПб.: Азбука, Азбука-Аттикус, 2012. – С. 62 – 78.
22. Информационный подход в междисциплинарной перспективе: материалы круглого стола // Вопросы философии. – 2010. – № 2. – С. 84 – 112.
23. Меркулов И. П. Когнитивная модель сознания // Эволюция. Мышление. Сознание. – М.: Канон+, 2004. – С. 35 – 64.
24. Смолян Г. Л. Информации теория // Новая философская энциклопедия. В 4 т. – М.: Мысль, 2010. – Т. 2. – С. 141 – 142.
25. Столяров Ю. Н. Документ как частный случай информации // Восемнадцатая Международная конференция «Крым-2011». «Библиотека и информационные ресурсы в современном мире науки, культуры, образования и бизнеса». – Судак, Автономная Республика Крым, Украина. 4–12 июня 2011 г. – URL: http://www.gpntb.ru/win/inter-events/crimea2011/disk/142.pdf. – 5 с.
26. Меркулов И. П. Введение: формирование когнитивных представлений в эпистемологии // Эволюция. Мышление. Сознание. – М.: Канон+, 2004. – С. 3 – 34.

Материал поступил в редакцию 25.02.13.

Сведения об авторе

ПОЛТАВСКАЯ Елена Игоревна – кандидат педагогических наук, зав. Отделом хранения Научной музыкальной библиотеки им. С. И. Танеева Московской государственной консерватории им. П. И. Чайковского.

e-mail: poltavskaya.elen@gmail.com

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 81’322.5

В.М. Гриняк, А.С. Девятисильный

Классификация движущихся объектов типа «надводный-воздушный» в лингвистических переменных*

Рассматривается проблема идентификации воздушных объектов современными СУДС. В основе предлагаемого подхода - оценка высоты наблюдаемого объекта по измерениям дальности и азимута. В дальнейшем полученные оценки обрабатываются нечеткой системой типа Мамдани, определяющей степень принадлежности объекта к классу воздушных. Описана конфигурация нечеткой системы, даются рекомендации по её обучению. С помощью компьютерного моделирования показана конструктивность предлагаемого подхода для типичных ситуаций.

Ключевые слова: *управление движением судов, воздушный объект, радар, измерение, высота объекта, нечеткая система типа Мамдани*

ВВЕДЕНИЕ

Задача заблаговременного распознавания опасно сближающихся судов (одна из центральных функций системы управления движением судов) оформилась в настоящее время как особый раздел науки об управлении [1, 2]. Методологической основой распознавания опасного сближения судов является оценка параметров траектории движения каждого судна (координат, скоростей и т.д.) и их экстраполяция. Если суда идентифицированы как опасно сближающиеся, система управления движением генерирует тревожный сигнал и рекомендации по изменению траектории движения.

Если в зоне ответственности системы управления движением судов (СУДС) наряду с надводными объектами присутствуют маловысотные низкоскоростные воздушные (вертолеты), то это может в корне исказить представление о навигационной обстановке. Суть проблемы состоит в том, что ошибочное заключение судоводителя или оператора СУДС о воздушной цели как о морской (когда их скорости движения сравнимы), способно привести к ложной тревоге и ошибочным управленческим решениям. Проблема частично решается применением автоматической идентификационной системы - АИС на воздушном объекте (информация АИС позволяет, в том числе, однозначно идентифицировать тип цели). Вместе с тем, транспондерами АИС оснащаются далеко не все воздушные объекты, допускающие полет над акваторией, что требует селекции воздушных объектов расширением навигационных функций систем, образуемых на основе двухкоординатных радаров.

В настоящей работе исследуется возможность создания на базе двухкоординатных радаров информационной системы, обеспечивающей достоверную классификацию объектов типа «надводный-воздушный» с использованием идей, положенных в основу обучаемых нечетких систем.

МОДЕЛЬНЫЕ ПРЕДСТАВЛЕНИЯ И ПОСТАНОВКА ЗАДАЧИ

Проблема трехкоординатного наблюдения воздушных объектов двухкоординатными измерителями неоднократно привлекала внимание исследователей [3-8]. Была показана принципиальная возможность (хотя и с ограниченным эффектом) решения трехкоординатной задачи при использовании одного двухкоординатного радара; продемонстрирован результат при переходе к многопозиционному наблюдению, когда используется система нескольких двухкоординатных радаров. В ряде работ [6, 8] доказана перспективность оценки координат объектов в сферической системе φ, λ, R – соответственно, географические широта, долгота и расстояние от центра Земли до объекта (с учетом пространственной локальности рассматриваемой задачи за модель поверхности Земли принимается сфера).

Особенностью внешнего наблюдения, осуществляемого с помощью радаров, является отсутствие непосредственного измерения сил и моментов, обусловливающих движение объекта. Поэтому при

* Работа выполнена в рамках Государственного задания высшим учебным заведениям в части проведения научно-исследовательских работ, проект № 7.2104.2011

описании эволюции координат наблюдаемых объектов традиционно обращаются к кинематическим моделям следующего полиномиального вида:

$$\varphi_{k+1} = \varphi_k + \sum_{i=1}^{n_\varphi} a_i^\varphi(k) T^i,$$
$$\lambda_{k+1} = \lambda_k + \sum_{i=1}^{n_\lambda} a_i^\lambda(k) T^i, \quad (1)$$
$$R_{k+1} = R_k + \sum_{i=1}^{n_R} a_i^R(k) T^i,$$
$$k = \overline{1,m},$$

где φ_k, λ_k, R_k – значения соответствующих координат объекта в момент времени t_k; n_φ, n_λ, n_R – максимальные значения степеней соответствующих полиномов; $a_i^\varphi(k)$, $a_i^\lambda(k)$, $a_i^R(k)$ – коэффициенты полиномов, отождествляемые со скоростями изменения соответствующих координат и функциями от их более старших производных; $T = t_{k+1} - t_k$; $t_k \in [t_1, t_m]$.

Информационная ситуация, обеспечиваемая сетью из L радаров, описывается моделью вида:

$$z_k^{(j)} = \begin{bmatrix} r^{(j)}(k) \\ \psi^{(j)}(k) \end{bmatrix} + \begin{bmatrix} \xi_r^{(j)}(k) \\ \xi_\psi^{(j)}(k) \end{bmatrix}, \quad (2)$$

где $z_k^{(j)}$ – вектор k-го измерения j-й станцией; $r^{(j)}(k)$ – дальность от объекта до j-й станции в момент времени $t_k^{(j)}$ (время k-го измерения j-й станцией); $\psi^{(j)}(k)$ – азимут объекта по отношению к j-й станции в момент времени $t_k^{(j)}$; $t_{k+1}^{(j)} - t_k^{(j)} = T^{(j)}$; $T^{(j)}$ – период вращения j-й станции; $\xi_r^{(j)}(k)$, $\xi_\psi^{(j)}(k)$ – инструментальные измерительные погрешности, причём $M[\xi_r^{(j)}(k)] = 0$, $M[\xi_r^{(j)}(k), \xi_r^{(i)}(m)] = D_r^{(j)} \delta_{ji} \delta_{km}$, $M[\xi_\psi^{(j)}(k)] = 0$, $M[\xi_\psi^{(j)}(k), \xi_\psi^{(i)}(m)] = D_\psi^{(j)} \delta_{ji} \delta_{km}$; $j = \overline{1,L}$; M[*] – оператор математического ожидания; δ_{ij} – символ Кронекера.

В соответствии с указанными модельными представлениями может быть поставлена обратная траекторная задача, описываемая уравнениями (1) и (2), ее цель – определение u-мерного вектора $s_k = (\phi_k, a_1^\phi(k), \ldots, a_{n_\phi}^\phi(k), \lambda_k, a_1^\lambda(k), \ldots, a_{n_\lambda}^\lambda(k), R_k, a_1^R(k), \ldots, a_{n_R}^R(k))^T$ по измерениям $z_k^{(j)}$, $j = \overline{1,L}$, $u = \dim s_k$.

МЕТОД РЕШЕНИЯ ЗАДАЧИ

Общим методом решения таких обратных задач является их линеаризация около некоторого опорного решения, характеризующего априорные представления о движении объекта. Допуская наличие опорного решения, будем говорить о сведении исходной задачи к задаче «в малом» с искомым вектором $\delta s_k = (\delta\phi_k, \delta a_1^\phi(k), \ldots, \delta a_{n_\phi}^\phi(k), \delta\lambda_k, \delta a_1^\lambda(k), \ldots, \delta a_{n_\lambda}^\lambda(k), \delta R_k, \delta a_1^R(k), \ldots, \delta a_{n_R}^R(k))^T$,

где δs_k – вектор погрешностей априорных представлений. Линеаризация исходных задач (1), (2) приводит её к следующему виду «состояние-измерение»:

$$\delta s_{k+1} = A_k \delta s_k + q_k,$$
$$\delta z_k^{(j)} = H_k \delta s_k + \xi_k^{(j)}, \quad (3)$$
$$j = \overline{1,L},$$

где q_k – вектор не моделируемых параметров движения, A, H – матричные коэффициенты (матрицы частных производных) с размерностью, соответственно, ($u \times u$) и ($2 \times u$). Преобразование уравнений (3) к конечномерному виду, характерному для задач метода наименьших квадратов, приводит исходную задачу к модели

$$\delta Z = \widetilde{H} \delta s_i + \widetilde{q}, \quad (4)$$

где δZ – полный вектор измерений на интервале наблюдения; δs_i – вектор погрешностей априорных представлений в момент времени t_i; $\widetilde{q}$ – вектор приведённых погрешностей измерений; $\widetilde{H}$ – матричный коэффициент размерности $N \times \dim s_i$, являющийся композицией матриц A и H; N – общее число обрабатываемых измерений (от всех станций).

Несмотря на то, что при $a_1^\varphi(i)$ и $a_1^\lambda(k)$, не равных одновременно нулю, система (4) не вырождена уже для одного радара ($L = 1$), а при наличии в системе нескольких радаров ($L > 1$) задача, в принципе, разрешима при любых возможных траекториях движения наблюдаемого объекта [5]. Для обеспечения практической разрешимости задачи необходимо ограничить размерность задачи (1), (2) так, чтобы движение объекта описывалось полиномами первой степени для угловых компонент и нулевой степенью для радиальной (т. е. $n_\varphi = 1$, $n_\lambda = 1$, $n_R = 0$, $s_i = (\varphi_i, a_1^\varphi(i), \lambda_i, a_1^\lambda(i), R_i)^T$). Это соответствует движению объектов на постоянной высоте без маневрирования на интервале наблюдения.

Характерным свойством рассматриваемой задачи (1), (2) является нерегулярность оценок радиальной координаты (т.е. высоты) маловысотных удалённых объектов, что связано с плохой обусловленностью системы (4), исходной нелинейностью задачи и конечной точностью измерений [6, 8]. Эта особенность задачи продемонстрирована на рис. 1, где приведена оценка высоты

надводного объекта (рис. 1а) и воздушных объектов, движущихся на высоте 100 м (рис. 1б) и 200 м (рис. 1в) – для случая двух РЛС, измеряющих дальность с погрешностью ±5 м, и азимут с погрешностью ±0.1°. Видно, что начиная с некоторого расстояния от системы радаров воздушный объект становится (по оценке высоты) неотличимым от морского: в данном случае это 5 км для объекта с высотой 100 м и 9 км – для объекта с высотой 200 м. Сами оценки высоты носят «изрезанный» характер со случайными выбросами. Такая картина является побудительным мотивом наряду с оцениванием собственно высоты объекта определять дополнительно ещё и «высотный класс» объекта, т. е. диапазон высот, которому принадлежит траектория объекта. В настоящей работе возможные диапазоны высот ограничены понятиями «морской» и «воздушный». При таком взгляде на проблему оказывается продуктивной идея формализации задачи в понятиях систем нечеткой логики.

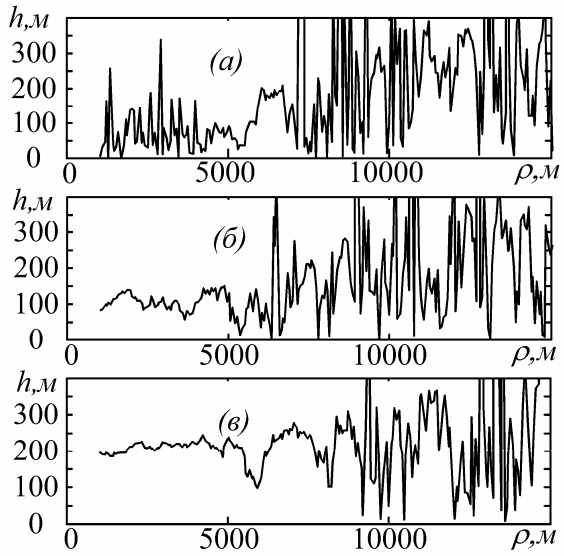

Рис. 1. Оценка высоты объекта по мере удаленияот радаров.
Здесь ρ - расстояние от системы радаров

Пусть $\hat{h}_i = \hat{R}_i - R_з$ – оценка высоты объекта над уровнем моря ($\hat{R}_i$ – оценка радиальной компоненты вектора s_i, $R_з$ – радиус Земли на уровне моря). С учетом особенности задачи будем считать, что основными информативными признаками, дающими представление о «высотном классе» объекта, являются оценка его высоты и сравнительный характер (степень «изрезанности», «нерегулярности») оценок высоты в различные моменты времени t_i. Введем лингвистическую переменную P_h «оценка высоты объекта» с термами «большая» и «малая» и функциями принадлежности типа «дополнение»:

$$\mu_{\text{малая}}(h) = 1 - \frac{1}{1+\exp(-a_h(h-c_h))},$$

$$\mu_{\text{большая}}(h) = \frac{1}{1+\exp(-a_h(h-c_h))}.$$

Пусть $\Delta_i = 2\left|\hat{h}_i - \hat{h}_{i-1}\right| \Big/ \left|\hat{h}_i + \hat{h}_{i-1}\right|$ – относительная разность между соседними оценками высоты. Введем лингвистическую переменную P_Δ «разность соседних оценок высоты объекта» с термами «большая» и «малая» и функциями принадлежности термов типа «дополнение»:

$$\lambda_{\text{малая}}(\Delta) = 1 - \frac{1}{1+\exp(-a_\Delta(\Delta - c_\Delta))},$$

$$\lambda_{\text{большая}}(\Delta) = \frac{1}{1+\exp(-a_\Delta(\Delta - c_\Delta))}.$$

Пусть u_i – степень принадлежности наблюдаемого объекта к диапазону высот «воздушный» в момент времени t_i. Введем лингвистическую переменную P_u «высотный диапазон объекта» с термами «надводный» и «воздушный» и функциями принадлежности типа «дополнение»:

$$\nu_{\text{надводный}}(u) = 1 - \frac{1}{1+\exp(-a_u(P - c_u))},$$

$$\nu_{\text{воздушный}}(u) = \frac{1}{1+\exp(-a_u(P - c_u))}.$$

Величины $\hat{h}_i$ и Δ_i (вход) обрабатываются нечеткой системой типа Мамдани M [9], показанной на рис. 2, на выходе которой формируется числовое значение u_i – степень принадлежности наблюдаемого объекта к диапазону высот «воздушный» в момент времени t_i. Машина нечеткого вывода работает согласно системе правил, представленной в таблице.

Система правил машины нечеткого вывода типа Мамдани

№	P_h	P_Δ	P_u
1	большая	большая	надводный
2	большая	малая	воздушный
3	малая	большая	надводный
4	малая	малая	надводный

Согласно этой системе правил, решение о том, что объект воздушный, принимается когда оценка высоты объекта достаточно велика, чтобы выделить его из морских и при этом она регулярна – относительная разность между соседними оценками незначительна. В противном случае принимается решение о том, что объект является надводным. Универсальное множество величины $\hat{h}_i$ целесообразно ограничить диапазоном [0, 100] м; если $\hat{h}_i > 100$, то этот вход принимается равным 100. Универсальное множество величины Δ_i ограничивается диапазоном [0, 1], если $\Delta_i > 1$, то этот вход принимается равным 1. Универсальное множество

выхода системы u_i есть отрезок [0, 1], причем $u_i = 0$ для однозначно надводных объектов и $u_i = 1$ для однозначно воздушных объектов.

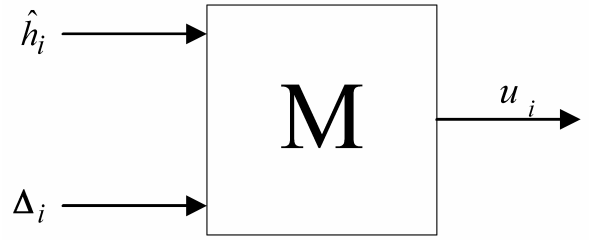

Рис. 2. Схема нечеткой системы, определяющей принадлежность объекта к диапазону высот «воздушный»

Настройка описанной системы состоит в задании количества измерений m от каждой радиолокационной станции (РЛС) и параметров функций принадлежности a_h, c_h, a_Δ, c_Δ, a_u, c_u.

ОБУЧЕНИЕ СИСТЕМЫ

Обучение системы (т. е. настройка параметров функций принадлежности a_h, c_h, a_Δ, c_Δ, a_u, c_u) может быть проведено с применением трёх различных стратегий.

Стратегия 1. Обучение полностью экспертным способом. В этом случае все коэффициенты назначаются экспертом.

Стратегия 2. Обучение на обучающей выборке с экспертным формированием заключений нечетких правил. В этом случае коэффициенты a_u, c_u назначаются экспертом, а коэффициенты a_h, c_h, a_Δ, c_Δ определяются настройкой системы на обучающей выборке.

Стратегия 3. Обучение полностью на обучающей выборке. В этом случае все коэффициенты системы определяются настройкой на обучающей выборке.

Обучающая выборка формируется следующим образом. Моделируется решение задачи (1), (2) при движении объекта на различных высотах, в том числе и при движении надводного объекта. В результате получаются оценки высоты объекта, подобные изображенным на рис. 1, формирующие входные данные обучающей выборки. Соответствующие им выходные данные обучающей выборки формируются экспертом: если характер оценки высоты объекта дает возможность отличить его от надводного, считается, что система выдаёт значение $u_i = 1$, и значение $u_i = 0$ – в противном случае. На рис. 3 показан пример формирования фрагмента такой обучающей выборки.

Здесь 3а – оценка высоты наблюдаемого объекта по мере его удаления от системы радаров (движение объекта моделируется на высоте 200 м), 3б – относительная разность соседних оценок высоты, 3в – решение эксперта о возможности выделить объект как воздушный: оценки высоты позволяют устойчиво сделать это до дальности, приблизительно, 10 км ($u_i = 1$), далее следует короткая зона, где объект не может быть выделен как воздушный, и поэтому отнесен к надводным ($u_i = 0$), затем до дальности, приблизительно, 11.5 км объект снова может быть отнесен к воздушным ($u_i = 1$), после чего следует сплошная зона, где он классифицируется как надводный ($u_i = 0$).

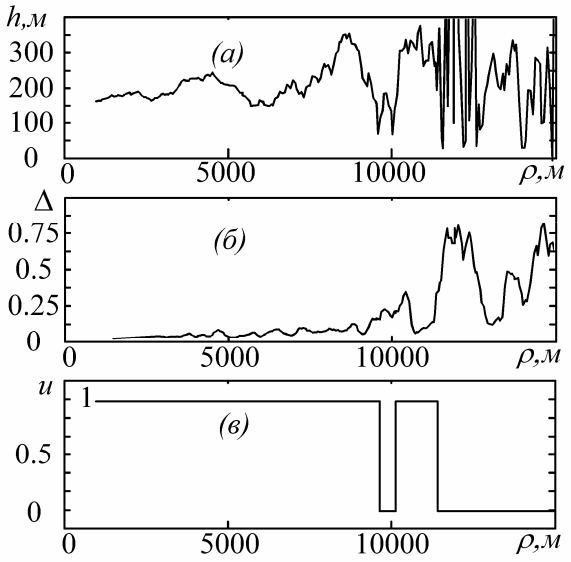

Рис. 3. Пример формирования обучающей выборки из оценок высоты, относительной разности соседних оценок высоты (вход) и степени принадлежности объекта к диапазону высот «воздушный» (выход).

Накапливая данные для различных высот движения объекта и множества возможных траекторий, формируют общую обучающую выборку, на базе которой и обучают нечеткую систему типа Мамдани (рис. 2) в рамках стратегии 2 или стратегии 3, пользуясь известными методами обучения систем такого типа [9, 10].

Результаты численного моделирования

При моделировании задачи было принято, что информационной базой СУДС являются два радара кругового обзора (например, типа Raytheon), находящихся на расстоянии 5 км друг от друга, с периодом обращения 3с и погрешностями измерений угла и дальности, соответственно, $\xi_\psi^{(j)}(k) \in [-0.1^\circ, 0.1^\circ]$, $\xi_r^{(j)}(k) \in [-5м, 5м]$. Количество измерений m от каждой станции было принято равным $m = 10$ и $m = 20$ (т. е. измерения набираются в течение 30 секунд и одной минуты).

Обучение системы проводилось в рамках стратегии 3, объём обучающей выборки составил около 10 000 значений «вход-выход», полученных при моделировании движения объекта по различным траекториям.

На рис. 4 показана траектория движения воздушного объекта, моделируемая для демонстрации решения задачи распознавания воздушных объектов с помощью предварительно обученной нечеткой системы типа Мамдани (см. рис. 2). Здесь I и II – радиолокационные станции, III – траектория объекта. Объект движется издалека по прямой со скоростью 20 м/с, приближаясь к РЛС; ρ – расстояние от объекта до линии, соединяющей радиолокационные станции.

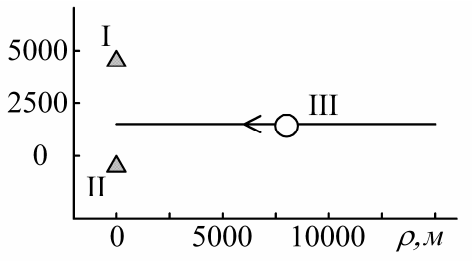

Рис. 4. Моделируемая конфигурация системы двух радаров и траектория движения объекта

На рис. 5 показаны результаты решения задачи оценки высоты объекта и оценки нечеткой системой его высотного диапазона. Здесь ρ – расстояние от объекта до линии, соединяющей радиолокационные станции, h – высота объекта, u – степень принадлежности объекта к диапазону высот «воздушный», P_u – высотный диапазон объекта. Задача моделировалась для объекта, движущегося на высоте 100 м с количеством измерений $m = 20$ (рис. 5а, 5в и 5д, левая колонка) и $m = 10$ (рис. 5б, 5г и 5е, правая колонка). Из рисунка видно, что, например, уверенное выделение воздушного объекта, движущегося на высоте 100 м, возможно до дальности ≈ 7000 м при $m = 20$ (рис. 5д) и до дальности ≈ 3000 м при $m = 10$ (рис. 5е). Такие дальности (по сути – границы применимости метода) вполне соответствуют размерам зон ответственности в акваториях морских портов, что указывает на пригодность предлагаемого метода селекции воздушных объектов для судоводительской практики.

ЗАКЛЮЧЕНИЕ

В настоящей статье обозначена проблема генерации ложных тревог при управлении коллективным движением судов, связанная с присутствием над акваторией воздушных объектов (вертолетов). Для корректной обработки системой управления движением судов таких объектов необходимо их идентифицировать. Алгоритм классификации объектов типа «надводный-воздушный» основан на вычислении высоты объекта по результатам измерений его дальности и азимута системой двухкоординатных радиолокационных станций и обработке полученных данных нечеткой системой типа Мамдани, предварительно обученной на моделируемых данных. Предлагаемый алгоритм позволяет принять решение о принадлежности объекта к классу надводных или воздушных объектов. В статье продемонстрированы границы применимости предлагаемой методики. В целом на основании анализа представленных данных можно сделать вывод о конструктивной разрешимости рассматриваемой задачи классификации. Результаты работы ориентированы на расширение навигационных функций современных систем управления движением судов.

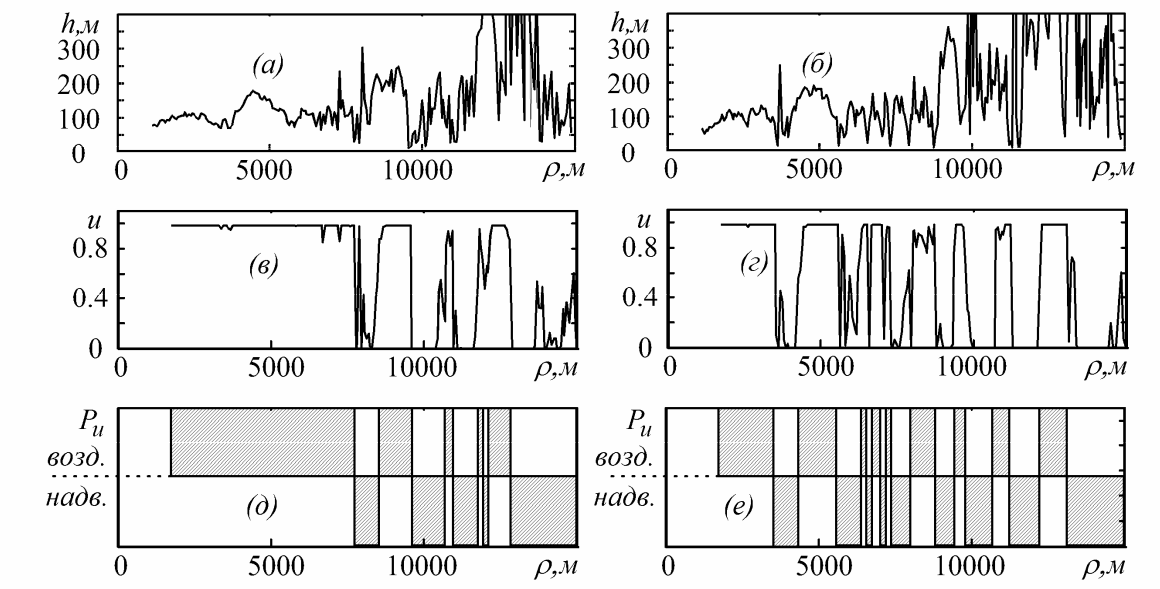

Рис. 5. Результат решения задачи

СПИСОК ЛИТЕРАТУРЫ

1. Tam Ch. K., Bucknall R., Greig A. Review of Collision Avoidance and Path Planning Methods for Ships in Close Range Encounters // The J. of Navigation. – 2009. – Vol.62, № 2. – P. 455-476.
2. Астреин В.В. Системы предупреждения столкновения судов, тенденции развития (к 40-летию МППСС-72) // Вестник Астраханского государственного технического университета. Серия: Морская техника и технология. – 2012. – №1. – С. 7-17.
3. Berle F.J. Multy radar tracking and multy sensor tracking in air defense systems // Electronic Technologies. – 1984. – Vol.28, № 4.
4. Nabaa N., Bishop R.H. Estimate Fusion for 2D Search Sensors // Proceedings of the AIAA Guidance, Navigation and Control Conference, August 1995, Monterey, CA.
5. Гриняк В.М. Исследование пространственной задачи навигации в условиях неполной измерительной информации // Дальневосточный математический журнал. – 2000. – Т.1, № 1. – С. 93-101.
6. Девятисильный А.С., Дорожко В.М., Гриняк В.М. Нейроподобные алгоритмы высотной классификации воздушных объектов // Информационные технологии. – 2001. – № 12. – С. 45-51.
7. Девятисильный А.С., Дорожко В.М., Гриняк В.М. Способ распознавания удалённых воздушных объектов: Патент №2206104 // Б.И. – 2003. – №16.
8. Гриняк В.М., Девятисильный А.С. Идентификация воздушных объектов в системах управления движением судов // Транспорт: наука, техника, управление. – 2012. – № 8. – С. 38-40.
9. Штовба С.Д. Проектирование нечетких систем средствами MatLab. – М.: Горячая линия телеком, 2007. – 288 с.
10. Nauk D., Klawonn F., Kruse R. Foundations of Neuro-Fuzzy Systems. – John Wiley & Sons, 1997. – 305 с.

Материал поступил в редакцию 02.04.13.

Сведения об авторах

ГРИНЯК Виктор Михайлович – кандидат технических наук, заведующий кафедрой Информационных систем и прикладной информатики Владивостокского государственного университета экономики и сервиса, кафедра.
e-mail: Viktor.Grinyak@vvsu.ru

ДЕВЯТИСИЛЬНЫЙ Александр Сергеевич – доктор технических наук, главный научный сотрудник Института автоматики и процессов управления ДВО РАН, Владивосток.
e-mail: devyatis@iacp.dvo.ru

УДК 81’32

И.Б. Яшков

Сравнение методов дискретизации и отбора непрерывных параметров для логико-комбинаторной классификации

Рассматриваются методы предобработки объектов, содержащих непрерывные атрибуты, для обучения алгоритма классификации на основе ДСМ-метода. Сравниваются методы дискретизации непрерывных параметров, не использующие информацию о распределении объектов по классам, с энтропийными, учитывающими метки классов при определении интервалов. Также рассматривается способ отбора атрибутов на основе энтропийной информации.

Ключевые слова: *дискретизация непрерывных параметров, атрибуция текстов, ДСМ-метод, энтропия, машинное обучение*

Многие методы машинного обучения предназначены для работы с объектами, все атрибуты которых являются дискретными, т.е. имеют область определения в виде конечного множества. Однако в реальной жизни исследователи часто сталкиваются с ситуацией, когда существенные для обучения параметры принимают значения из непрерывной области определения, как, например, множество положительных чисел, множество действительных чисел или доля в процентах. Один из подходов к анализу таких данных рассматривается в работе [1] на примере ДСМ-метода автоматического порождения гипотез (АПГ).

В своей работе мы также решили выбрать ДСМ-метод в качестве рассматриваемого метода интеллектуального анализа данных, что продиктовано несколькими причинами. Во-первых, очевидно, ДСМ-метод предназначен для работы с исключительно дискретными свойствами. Во-вторых, ДСМ-метод требователен к количеству свойств и их распределению на объектах: в работе [2] показано, что сложность алгоритма Норриса, используемого для генерации гипотез в стандартной реализации ДСМ-систем, линейно зависит от количества признаков и от количества генерируемых формальных понятий. В-третьих, неадекватная дискретизация может снизить абдуктивную способность порожденных гипотез.

Таким образом, задача отбора атрибутов и дискретизации тех из них, которые являются непрерывными, представляет собой одну из ключевых проблем этапа предобработки данных перед применением ДСМ-метода. Цель настоящей работы – сравнить некоторые из способов решения этой задачи и оценить их влияние на качество и скорость порождения гипотез при различных объемах обучающей выборки.

ПРЕДМЕТНАЯ ОБЛАСТЬ

В качестве тестовой задачи было выбрано автоматическое определение авторства текста (задача атрибуции текста). Это решение продиктовано рядом причин.

Во-первых, компьютерная лингвистика является слабо формализованной областью, но при этом некоторые задачи, которые традиционно решаются в ее рамках, работают с хорошо структурированными моделями. Определение авторства – одна из таких задач, это подтверждается, например, успешной практикой применения математических методов в судебном автороведении. ДСМ-метод положительно зарекомендовал себя именно для решения задач интеллектуального анализа хорошо структурированных данных в слабо формализованных областях [3].

Во-вторых, при любой задаче классификации текста исследователь-лингвист неизбежно сталкивается с проблемой выбора значимых атрибутов. Особенно это характерно для задачи определения авторства: для некоторых авторов существенны лексические характеристики речи (т.е. употребление конкретных слов), для других – синтаксические, такие как употребление союза «и» в начале предложения или частое употребление вводных конструкций. В некоторых исследованиях (например, [4]) рассматриваются смысловые характеристики: какие семантические поля наиболее широко представлены в текстах автора. Так как сокращение признакового пространства являлось одной из целей препроцессинга в нашем исследовании, это стало еще одним аргументом в пользу выбора задачи атрибуции.

Наконец, в-третьих, параметры текста в задачах классификации в большинстве своем представляют собой непрерывные величины, для которых нет универсальной шкалы. Даже лексические характеристики, которые можно было бы представлять бинарными признаками «встречается/не встречается» для каждой лексемы, для некоторых целей удобно иметь в виде вектора чисел из отрезка [0; 1]. Семантика этих весов может быть разной: при классификации новостей это «информативность» лексемы, а в задаче атрибуции – «характерность» слова для автора. Так или иначе, в отсутствие универсальных шкал, которые позволили бы оценить каждый параметр, – принял ли он «низ-

кое», «высокое» или «среднее» значение – встает проблема дискретизации.

С учетом всего этого, задача атрибуции кажется хорошо подходящей для поставленных целей. Качество порожденных гипотез для классификации текстов полностью определяется качеством предварительной обработки данных, так как в универсуме атрибутов текста, изначально определенных аналитиком, с высокой долей вероятности содержатся несущественные, а их значения подлежат разбиению на интервалы, для которых заранее неизвестна шкала оценки. Метрики качества и скорости итоговой классификации, таким образом, в этой задаче одновременно служат оценкой адекватности предобработки исходных данных.

ЛИНГВИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ

Текстами, на которых обучался и тестировался алгоритм классификации, в данной задаче послужили стихотворения, разбитые на четыре класса по их авторам. Всего было выбрано четыре автора: М.Ю. Лермонтов, М.И. Цветаева, О.Э. Мандельштам, И.А. Бродский.

Тексты были разбиты на токены (слова, знаки препинания или неделимые последовательности символов) и сегментированы (разбиты на предложения). Над последовательностями токенов был проведен морфологический анализ со снятием частеречной и грамматической омонимии. Таким образом, после лингвистической обработки текст содержал множество предложений S, каждое из которых представлялось множеством словоформ W и типом последнего пунктуационного токена: точка/восклицательный знак/вопросительный знак. Словоформа в свою очередь определялась лексемой, к парадигме которой она принадлежит, частью речи и набором пар *<грамматический признак, значение>*.

Используя данное представление, для каждого текста мы посчитали следующие характеристики:

1. **Длина текста** в символах.
2. **Средняя длина словоупотреблений** в тексте.
3. **Среднее количество словоупотреблений в предложении**.
4. **Доля существительных** среди всех словоупотреблений.
5. **Доля финитных глаголов** среди всех словоупотреблений.
6. **Доля инфинитивов** среди всех словоупотреблений.
7. **Доля полных прилагательных** среди всех словоупотреблений.
8. **Доля полных причастий** среди всех словоупотреблений.
9. **Доля кратких прилагательных и причастий** среди всех словоупотреблений.
10. **Доля наречий** среди всех словоупотреблений.
11. **Доля деепричастий** среди всех словоупотреблений.
12. **Доля числительных** среди всех словоупотреблений.
13. **Доля личных местоимений** среди всех словоупотреблений.
14. **Доля частиц** среди всех словоупотреблений.
15. **Доля предлогов** среди всех словоупотреблений.
16. **Доля союзов** среди всех словоупотреблений.
17. **Доля императивов** среди всех словоупотреблений.
18. **Доля уменьшительно-ласкательных существительных** среди всех существительных.
19. **Доля местоимений «я» и «мой»** среди всех словоупотреблений.
20. **Доля местоимений «мы» и «наш»** среди всех словоупотреблений.
21. **Доля финитных глаголов в первом лице** среди всех глаголов.
22. **Доля неизвестных лексем** среди всех словоупотреблений. Неизвестными лексемами считаются такие, которые не встретились в словаре, используемом при морфологическом анализе (в данной задаче использовался словарь на 100 000 лексем).
23. **Словарный запас.** Доля уникальных лексем среди всех словоупотреблений.
24. **Доля чисел, записанных цифрами,** среди всех словоупотреблений.
25. **Доля не-кириллических символов**.
26. **Доля вопросительных предложений**.
27. **Доля восклицательных предложений**.
28. **Индекс читабельности Флеша**.

Формула, предложенная Рудольфом Флешем, для анализа сложности восприятия текста. Использование формулы Флеша в задаче атрибуции текста рассматривается в работе [5], при этом формула приводится в модификации для русского языка, предложенной И.В.Оборневой [6]:

$$K = 206{,}836 - 65{,}14 \times W - 1{,}52 \times S\,,$$

где W – средняя длина слогов, S – средняя длина предложений в словах.

K выражается в значениях от 0 до 100. Оценка 100 говорит о том, что человек с минимальным уровнем образования сможет ответить на три четверти вопросов по тексту; оценка 0 – текст доступен лишь узким специалистам; оценка 60 – текст стандартный по трудности для среднего читателя.

Проводились также эксперименты с включением в список признаков лексических параметров (частот использования конкретных лексем), однако при этом возникала дополнительная задача предварительного отбора лексем для включения в пространство признаков перед дискретизацией. Несмотря на повышение абсолютной точности атрибуции при наличии этих признаков, процесс предварительного отбора вносил дополнительную неопределенность в сравнительный анализ способов дискретизации. Так как мы не ставили перед собой цель достичь максимальной точности атрибуции, а оценивали только относительное изменение качества и скорости, от включения лексических признаков в итоге было решено отказаться (ниже будет показано, что точность атрибуции тем не менее сравнима с точностью современных алгоритмов).

ВЫБОР МЕТОДОВ ДИСКРЕТИЗАЦИИ ПАРАМЕТРОВ

Очевидно, что список признаков полностью состоит из непрерывных параметров. Из них ни один, кроме индекса Флеша, не имеет универсальной шкалы оценки, по которой можно было бы определить, является ли его значение стандартным или отклоняется от него в ту или иную сторону. Следовательно, мы не могли воспользоваться способами дискретизации, обусловленными предметной областью (ср., например, с работой [1], в которой для уровня канцерогенности соединений существует заранее заданная шкала).

В прошлом нам приходилось сталкиваться с задачей, в которой требовалось провести классификацию с помощью ДСМ-метода и объекты содержали непрерывные параметры. Опыт решения этой задачи описан в работе [7] – одним из этапов анализа в ней является разбиение пациентов на группы по схожести признаков. Несмотря на то, что проблема дискретизации в работе [7] не рассматривается, одним из сделанных нами выводов было соображение, что качество разбиения может повыситься, если выбрать другой метод разбиения исходных непрерывных параметров на интервалы. В частности, у экспертов в предметной области состав классов вызывал сомнения.

Для дискретизации признаков в работе [7] использовался метод разбиения на равные интервалы, при этом количество интервалов зависело от количество объектов N и выбиралось по правилу Стёрджесса равным $1+\log_2(N)$ [8]. Это правило дает сравнительно большое количество интервалов (для 100 объектов в выборке, например, это число уже равно 8), поэтому нам было интересно выбрать для сравнения дискретизацию с относительно небольшим количеством интервалов. Из этих соображений мы проводили также разбиение на два и три интервала.

Три этих дискретизатора (два интервала, три интервала, $1+\log_2(N)$ интервалов) не используют информацию о метках классов у объектов обучающей выборки. Здравый смысл подсказывает, что учет распределения объектов по классам должен давать более адекватное разбиение. Это предположение рассматривается в работе [9], авторы которой показывают, что так называемые дискретизация с учителем (т. е. разбиение на интервалы, учитывающее метки классов) увеличивает точность наивного байесовского классификатора и классификатора на основе дерева решений, построенного с помощью алгоритма C4.5.

Из рассмотренных в [9] методов разбиения наибольший прирост точности классификации по сравнению с разбиением без учителя дал метод рекурсивной энтропийной дискретизации. Этот метод использует правило минимальной энтропии, описанное в работах [10] и [11].

Для начала приведем определение энтропии (степени неопределенности) для множества объектов обучающей выборки S:

$$Ent(S) = -\sum_{x \in X} p(x) \log_2 p(x) \ ,$$

где X – множество классов в S, p(x) – отношение мощности класса x к мощности S. Когда Ent(S) = 0, множество S является идеально классифицированным, т. е. все его объекты принадлежат одному классу.

Энтропией разбиения множества S по границе T признака A называется величина

$$E(A,T;S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2) \ ,$$

где S_1 – множество объектов, у которых значение признака A меньше T, а S_2 – множество объектов, у которых значение признака A больше T.

Метод рекурсивной энтропийной дискретизации состоит в следующем: для данного признака A и множества S находится граница T_{min}, при которой энтропия разбиения S – минимальна. T_{min} выбирается в качестве границы интервала, после чего метод рекурсивно применяется к множествам S_1 и S_2. В качестве условия остановки рекурсии в работе [11] рассматривается следующее: разбиение прекращается, когда выполнено условие:

$$Ent(S) - E(A,T;S) < \frac{\log_2(N-1)}{N} + \frac{\Delta(A,T;S)}{N} \quad (*) \ ,$$

где

$$\Delta(A,T;S) = \log_2(3k-2) - \left[\begin{array}{l} |X| \times Ent(S) - |X_1| \times Ent(S_1) - \\ -|X_2| \times Ent(S_2) \end{array} \right] \ ,$$

где X_i – множество классов в S_i.

Это условие можно трактовать по-разному: либо а) разбиение не производится, как только условие (*) выполнено, либо б) разбиение не производится на следующем шаге. Мы решили проверить обе трактовки. Заметим, что в случае (а), если на первом же шаге условие (*) будет выполнено, множество границ интервалов останется пустым, что равносильно исключению признака из классификации. Мы также решили попробовать промежуточный вариант между (а) и (б) – если условие (*) выполнено на первом шаге, делим по T_{min} на два интервала, в противном случае останавливаем разбиение.

Таким образом, у нас получился набор из шести методов предобработки:

1. **EL2** (*equal length 2*). Разбиение на два равных интервала.
2. **EL3** (*equal length 3*). Разбиение на три равных интервала.
3. **ELS** (*equal length Sturges*). Разбиение на $1+\log_2(N)$ равных интервалов.
4. **ENA** (*entropy a*). Рекурсивная энтропийная дискретизация с остановкой непосредственно при выполнении условия (*).
5. **ENAB** (*entropy ab*). Рекурсивная энтропийная дискретизация с остановкой непосредственно при выполнении условия (*) на любом шаге, кроме первого.
6. **ENB** (*entropy b*). Рекурсивная энтропийная дискретизация с остановкой на следующем шаге после выполнения условия (*).

Заметим, что ENA, кроме задачи дискретизации, также решает поставленную нами задачу отбора признаков.

РЕЗУЛЬТАТЫ СРАВНЕНИЯ

Для сравнения методов было решено постепенно увеличивать обучающий корпус с 10 до 80 объектов в классе, т.е. с 40 до 320 объектов при сбалансированной выборке. Для каждого объема и каждого дискретизатора мы запускали ДСМ-метод автоматического порождения гипотез.

Так как в данном случае нас интересовали гипотезы только о принадлежности объектов к классу, для каждого из классов с помощью алгоритма Норриса строилась решетка только таких понятий, в которых присутствовал признак класса, т. е. объекты из различных классов априори считались различными. Таким образом, сходство определялось как пустое множество в случае несовпадающих классов, либо как пересечение множеств значений свойств, если классы совпадали.

Следующим шагом из каждой решетки отбирались только такие гипотезы H, которые были порождены не менее чем H_{min} объектами при наличии не более H_{max} контрпримеров (т.е. таких объектов о принадлежности к другим классам, которые включали бы множество свойств из H). H_{min} и H_{max} являются параметрами алгоритма классификации, которые в данной задаче были зафиксированы в значениях 5 и 2 соответственно.

Наконец, из оставшихся гипотез выбиралось финальное множество максимальных по включению. Эти гипотезы после каждого запуска ДСМ-метода тестировались на множестве объектов объемом 400 (по 100 объектов в каждом классе). Точность классификации в процентах в итоге принималась за метрику качества классификации.

На рис. 1 представлено сравнение качества классификации при всех возможных комбинациях объема обучающей выборки и типа дискретизации:

Дополнительно оценивалась скорость классификации. Чтобы избежать зависимости от конкретной реализации алгоритмов построения решетки понятий и отбора гипотез, воспользуемся оценкой сложности алгоритма Норриса, приведенной в работе [2]. Его сложность квадратично зависит от количества объектов в обучающей выборке |G|, и линейно – от количества признаков |M| и мощности решетки |L|. На рис. 2 приведены зависимости величины $|G|^2 \times |M| \times |L|$ от количества объектов в классе для различных типов дискретизации:

Для справки приведем также на рис. 3 график изменения количества свойств у объектов для различных типов дискретизации и объемов обучающей выборки:

Как можно видеть из приведенных рисунков, при малом объеме обучающей выборки энтропийные алгоритмы дискретизации значительно проигрывают разбиению на равные интервалы, из которых лучше всего себя показывает EL2 (разбиение на два равных интервала). Однако с ростом количества объектов, качество классификации при энтропийной дискретизации также растет и в финальном эксперименте с 80 объектами в каждом классе обучающей выборки алгоритм ENA в итоге побеждают остальные, в то время как качество классификации с разбиением на равные интервалы при увеличении объема выборки постепенно снижается.

В итоге, если сравнивать ENA с его основным «конкурентом» EL2, можно видеть, что в последнем эксперименте он дает лучшую точность классификации (61% против 59%), работая на порядок быстрее, и обходится наименьшим количеством признаков, что всегда повышает понятность порожденных гипотез для экспертов. Заметим также, что итоговое качество классификации вполне сравнимо с результатами, приводимыми для алгоритмов атрибуции коротких текстов [12].

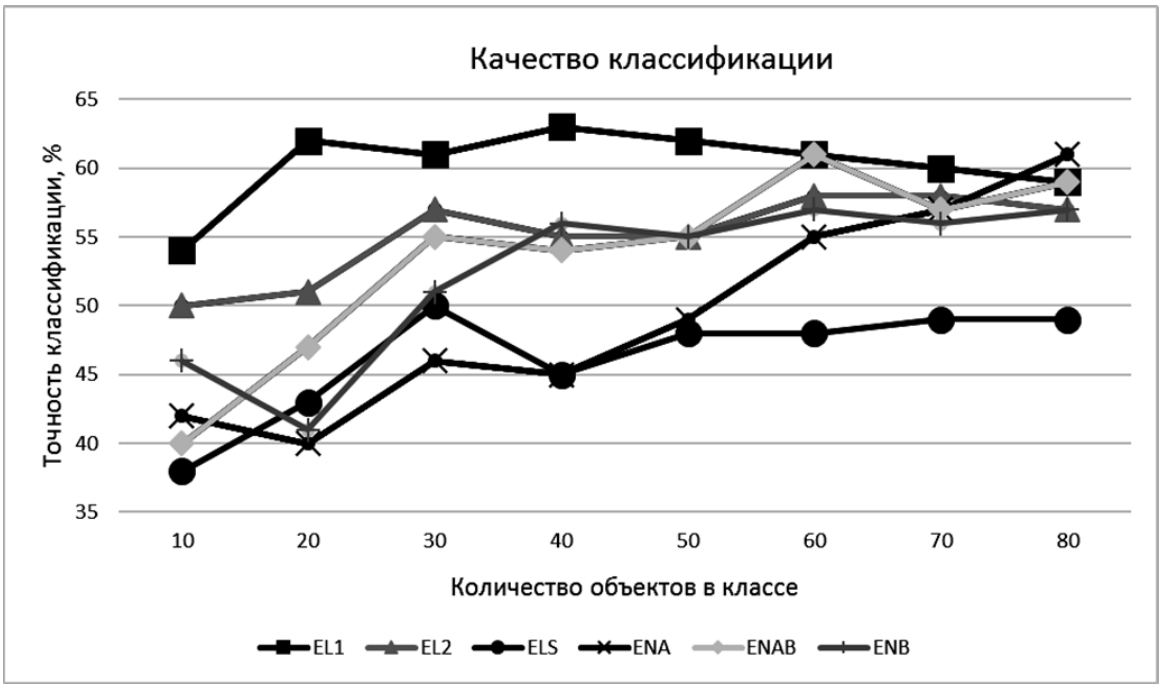

Рис. 1. Качество классификации

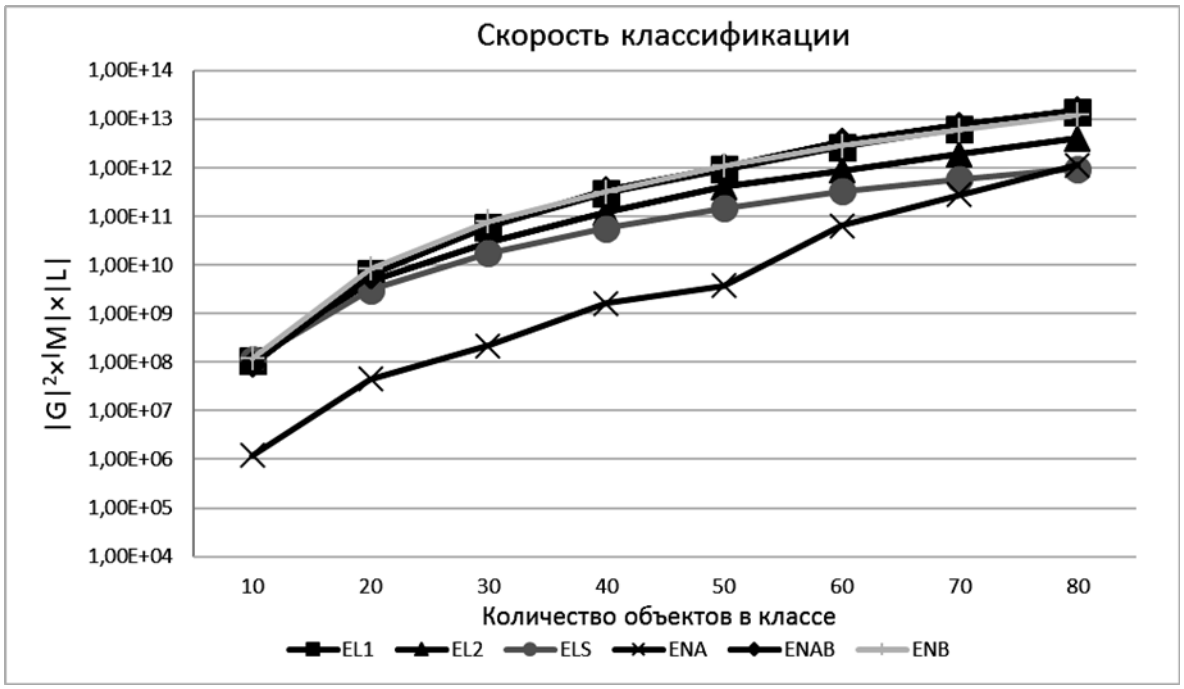

Рис. 2. Скорость классификации

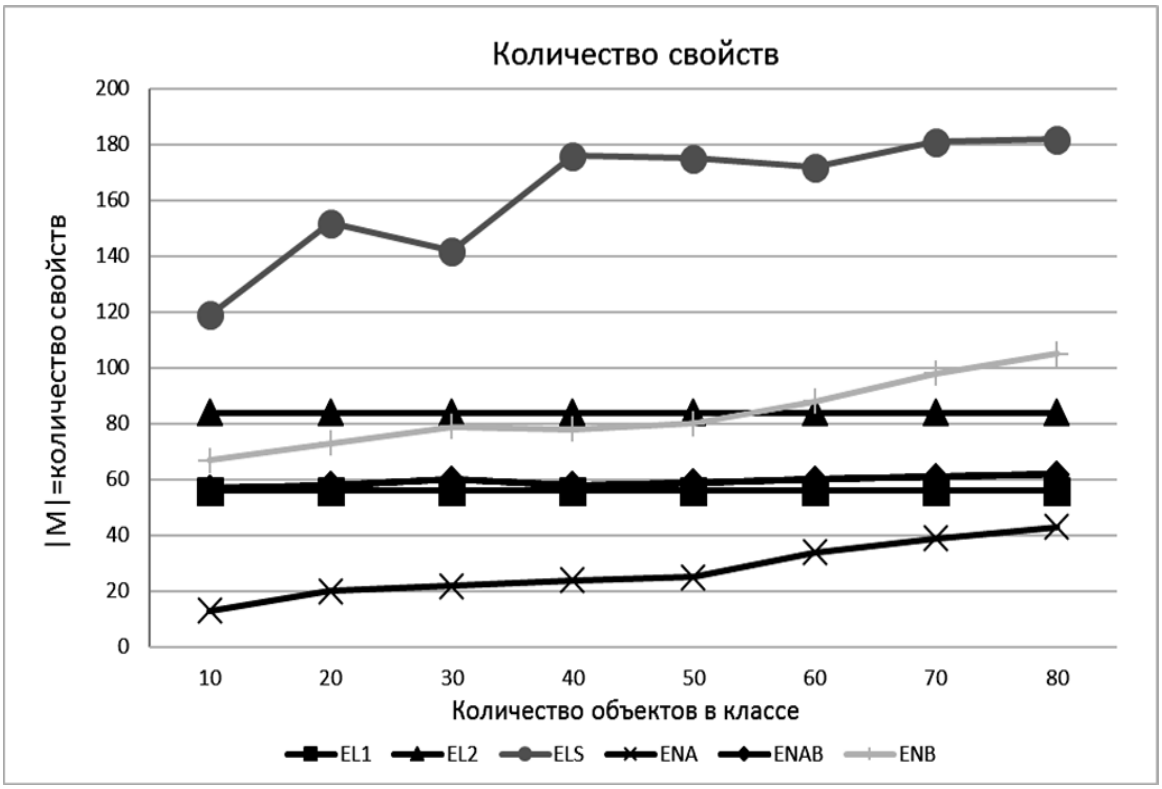

Рис. 3. Количество свойств

Основной вывод проведенного нами сравнения следующий: при достаточном объеме обучающей выборки рекурсивная энтропийная дискретизация с отбором признаков на первом шаге действительно позволяет повысить точность и скорость классификации, однако при малом объеме лучше обойтись дискретизацией без учителя с наименьшим количеством интервалов, а именно – двумя для каждого непрерывного признака.

НАПРАВЛЕНИЯ ДАЛЬНЕЙШЕЙ РАБОТЫ

Первый вопрос, возникающий после прочтения сформулированного вывода: как определить достаточный объем выборки, при котором будет целесообразно применять энтропийную дискретизацию? Дальнейшая наша работа будет направлена на выявление такого критерия. Кроме того, проведенные эксперименты ставят новые задачи, которые планируется решить в будущем.

Для получения более точных данных о качестве и скорости классификации, процесс порождения гипотез для каждой комбинации объема выборки и дискретизатора нужно проводить несколько раз, каждый раз рандомизируя обучающее множество, и высчитывать усредненные показатели и дисперсию.

Планируется также протестировать другие методы дискретизации без учителя (например, разбиение на интервалы, включающие в себя одинаковое количество объектов) и с учителем и их комбинации: например, дискретизация без учителя с энтропийным отбором.

Не исключена возможность того, что к разным признакам потребуется разный подход: для этого мы попробуем расширить признаковое пространство атрибутами другой природы, а именно – лексическими и семантическими.

Кроме задачи атрибуции, мы планируем сравнить различные подходы к дискретизации в других задачах классификации текстов (рубрикация, определение стиля, определение искусственности/естественности), а также в задачах, не имеющих отношения к компьютерной лингвистике, чтобы иметь более репрезентативную выборку сценариев.

Автор статьи ставит перед собой цель с помощью решения поставленных задач и сравнения метрик классификации определить универсальную процедуру отбора и дискретизации непрерывных параметров, которая помогла бы повысить точность и скорость логико-комбинаторной классификации, а также сделать порождаемые ДСМ-гипотезы более понятными для экспертов предметной области.

СПИСОК ЛИТЕРАТУРЫ

1. Максин М.В. Об одном подходе к проблеме комбинированного использования логических и численных методов в интеллектуальном анализе данных // Научно-техническая информация. Сер. 2. – 2004. – № 10. – С. 14–19.
2. Kuznetsov S.O., Obiedkov S.A. Comparing Performance of Algorithms for Generating Concept Lattices // Journal of Experimental and Theoretical Artificial Intelligence – 2002. – Vol. 14. – P. 189–216.
3. Финн В.К. Об особенностях ДСМ-метода как средства интеллектуального анализа данных // Научно-техническая информация. Сер. 2. – 2001. – № 5. – С. 1–3.
4. Хоменко А.Ю., Романова Т.В. Апробация методов математической статистики при атрибуции текста в рамках судебного автороведения // «Филология, искусствоведение и культурология в XXI веке»: материалы международной заочной научно-практической конференции – Новосибирск: СибАК, 2013. – С. 64–75.
5. Мальковский М.Г., Ширшова К.П. Методы и алгоритмы определения личностных характеристик автора текста // Сборник научных трудов по материалам международной научно-практической конференции «Перспективные инновации в науке, образовании, производстве и транспорте - 2010». Том 4. Технические науки. – Одесса: Черноморье, 2010. – С. 14–16.
6. Оборнева И.В. Математическая модель оценки учебных текстов // Вестник Московского городского педагогического университета. Серия: Информатика и информатизация образования. – 2005. – № 4. – С. 152–158.
7. Высочин И.В., Яшков И.Б., Виноградов Д.В. Построение модели мобилизации и афереза кроветворных стволовых клеток у пациентов с онкологическими заболеваниями с использованием гибридной ДСМ-системы // Эфферентная и физико-химическая медицина – 2012. – № 1. – С. 51–54.
8. Sturges H. The choice of a class-interval // Journal of the American Statistical Association. – 1926. – Vol. 21, № 153. – P. 65–66.
9. Dougherty J., Kohavi R., Sahami M. Supervised and Unsupervised Discretization of Continuous Features // Proceedings of the 12th International Conference on Machine Learning (ICML-1995). – San Francisco: Morgan Kaufmann, 1995. – P. 194–202.
10. Catlett J. On changing continuous attributes into ordered discrete attributes // Proceedings of the European Working Session on Learning. – Berlin: Springer-Verlag, 1991. – P. 164–178.
11. Fayyad U.M., Irani K.B. Multi-interval discretization of continuous-valued attributes for classification learning // Proceedings of the 13th International Joint Conference on Artificial Intelligence. – San Francisco: Morgan Kaufmann, 1993. – P. 1022–1027.
12. Романов А.С., Мещеряков Р.В. Идентификация авторства коротких текстов методами машинного обучения // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог-2010». Вып. 9 (16). – М.: Изд-во РГГУ, 2010. – С. 407–413

Материал поступил в редакцию 28.05.13.

Сведения об авторе

ЯШКОВ Игорь Борисович - аспирант ВИНИТИ РАН, Москва,
e-mail: IgorYashkov@gmail.com

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322

А.В. Алексеев

К вопросу о символическом семиозисе слова в истории языка

Рассматриваются способы реализации символических структур в лексической семантике. Символ понимается как механизм передачи информации, при котором возникает сверхбинарная знаковая структура: означаемое первого уровня одновременно является означающим для означаемого второго уровня. Выявить символические формы в лексике возможно лишь при диахроническом подходе. В истории слова проявлением символа становятся первоначально внутренняя форма, а впоследствии лексические значения особого типа (диффузные, символические). В процессе символического семиозиса также формируются культурные коннотации слова, посредством которых слово входит в систему культурных знаков.

Ключевые слова: *символическое значение, лексическая диффузность, внутренняя форма слова, культурная коннотация*

Ориентируясь на семиотические традиции филологии и философии двадцатого века, мы признаем символ ведущим средством информационного взаимодействия. В ряде работ символический семиозис может противопоставляться семиозису собственно знаковому, и таким образом выделяются два принципиально разных механизма формирования и передачи информации, эти механизмы могут определяться как «означивание и символизация» [1, с. 51]. В других случаях символ признается высшей формой знака, «конденсатором» знаковости [2, с. 44], однако, отмечается, что символ реализует также и механизмы, противоположные знаковым, а именно образные. В случае признания универсального характера символа, он определяется как своеобразная «точка равновесия» между знаковостью и образностью [3, с. 114]. Таким образом, можно выделить два подхода к пониманию символа: 1) символ и знак – два способа информационного взаимодействия; 2) символ объединяет качества знака и образа.

Следует заметить, что указанные трактовки сходным образом описывают проблематику символического семиозиса – различия могут быть сведены к своеобразному терминологическому сдвигу: при втором подходе образность отождествляется с символизмом в том смысле, в каком последний понимается как особый информационный механизм в первом подходе. Иными словами, «знак + символ = семиозис» при первом подходе и «знак + образ = символический семиозис» – при втором подходе.

В настоящей статье нам придется учитывать дихотомию знака и символа (первый подход), поскольку именно такая дихотомия содержится в основополагающих для современной лингвистической семиотики работах Ф. де Соссюра. Мы, однако, вопреки Ф. де Соссюру, который признавал языковые единицы, в частности слова, знаками, но не символами, предполагаем выявить символические качества слов, и интересовать нас будет в первую очередь исторический (диахронический) аспект. При структуралистском подходе, который базируется на работах Ф. де Соссюра, в качестве важнейших, основных свойств языкового знака выделяются его условность и конвенциональность. На фоне этого символ должен определяться как семиотическая структура, содержащая мотивированную связь между означающим и означаемым, в наиболее радикальных концепциях подобная связь определяется как закономерная и «внутренне-обязательная» [4, с. 196].

В рамках структурализма понимаемый подобным образом символ невозможен в естественном языке: Ф. де Соссюр признавал существование в языке лишь «независимых символов» (иными словами, знаков), основным свойством которых должно быть «отсутствие всякого рода видимой связи с обозначаемым объектом» [5, с. 91]. Обычный символ сохраняет естественную связь между означающим и означаемым, мотивированность, которая Ф. де Соссюру представлялась «рудиментом» [6, с. 70]. То, что мы понимаем как важнейшее средство сохранения и передачи информации, Ф. де Соссюром рассматривалось как побочное, случайное, не актуальное для функционирования языковой единицы качество. По замечанию Ц. Тодорова, «в основном Соссюр упоминает символы лишь для того, чтобы отрицать их...» [1, с. 339].

Структуралистское положение о произвольности языкового знака оказалось актуальным и для более поздних лингвистических направлений, в частности, «…в основу прагматики 1950-х гг. был положен тезис об отсутствии «естественной» связи между «оз-

начаемым» и «означающим» как двумя сторонами знака...» [7, с. 33]. Это положение признается и рядом современных лингвистов [8, с. 227]. Таким образом, в науке широко распространено мнение о противопоставленности символа и языкового знака за счет преобладания у последнего произвольности и конвенциональности [9, с. 127].

Нам, однако, представляется, что отрицание Ф. де Соссюром символичности языка было чрезмерно категоричным и вело к обеднению лингвистической семиотики, так как не позволяло изучать символические аспекты языка в целом [1, с. 342] и символические качества слова как центральной языковой единицы. Современные исследователи отмечают, что «Ф. де Соссюр… во многом дискредитировал это понятие» <символ – *А.А.*>, поскольку из его тезисов следует, что лингвистика может изучать лишь искусственные знаки [10, с. 1]. Вместе с тем и прежде отмечалось, что «…иконические и индексальные составляющие языкового знака слишком часто недооценивались…» [11, с. 125]. Нам представляется, что признание полной произвольности языкового знака – явное преувеличение, возникающее из-за непонимания исторического характера языковой коммуникации. Историчностью языковых единиц обеспечивается совмещение в них качеств символичности (образности, мотивированности) и чистой знаковости (произвольности).

Под историчностью следует понимать прежде всего обусловленность наличной языковой системы ее предшествующим состоянием, следовательно, обусловленность выбора тех или иных новых знаковых средств набором уже существующих языковых знаков. Это то качество, которое, как мы полагаем, может быть названо внутренней формой языка в широком смысле этого термина. Историчность – это неизбежное изменение качеств каждого отдельно взятого языкового знака, в частности, слова, которое, будучи символом (в интегральном, универсальном смысле), характеризуется первоначальной образностью и последующей знаковостью. При рождении слова его мотивированность (прямая звукоподражательная либо опосредованная внутренней формой) максимальна. По мере функционирования слова в качестве языкового символа усиливаются его знаковые качества, конвенциональность. Подобная динамика соответствует закономерностям развития символических форм, у которых по мере их развития «иконические свойства ослабевают, а знаковые свойства... усиливаются» [10, с. 7]. Однако полная произвольность все же оказывается при этом фикцией: наблюдаемое в синхронном описании отсутствие связи между формой и содержанием языкового знака должно объясняться тем, «что эта связь стерта, демотивирована» [12, с. 25].

Признание слова символом – важнейшее условие его плодотворного изучения в контексте исторической лексикологии и концептологии. В.В. Колесов определяет символ как содержательную форму слова [13, с. 53]. Слово является не единственной языковой единицей, проявляющей символичность, но именно в слове качества символичности реализуются наиболее ярко. Именно вследствие своей символичности слова могут передавать смысл, «способны образовывать суждения» [11, с. 125].

Следовательно, в слове мы должны обнаружить динамическое объединение образности (символичности) и знаковости, выявляемое при диахроническом подходе. Каковы же конкретные проявления символичности слова в его истории? Как указывалось, в момент появления слово характеризуется звукоподражательной мотивированностью или внутренней формой. Звукоподражательные слова (ономатопеи), несмотря на обусловленность связи между означающим и означаемым, нельзя считать символическими формами, так как они не обладают необходимыми структурными особенностями символа. Минимальная структура символа образуется удвоением бинарной связи означаемого и означающего, при котором означаемое первого уровня одновременно является означающим второго уровня: означающее 1 <–> (означаемое 1 = означающее 2) <–> означаемое 2. Подобное формирование в символе вторичного смысла отмечалось большинством исследователей символических форм, при этом «означаемое 2» могло рассматриваться как культурно ценное содержание [2, с. 240], иносказательный смысл [14, с. 44], подсознательная информация [15, с. 16] и т.д. Такой двуплановости смысла нет у ономатопей, хотя само явление звукоподражания соотносимо с фонетическим символизмом – однако теория звукосимволизма требует дополнительного обсуждения, невозможного в данной работе.

Слова, не непосредственно фонетически мотивированные, а обладающие внутренней формой, несомненно, соответствуют троичной символической модели. Подобные семантические структуры будут нами выявлены в описании А.А. Потебни, а затем рассмотрены на примере нескольких существительных, называющих социальные и духовные явления русской культуры в ее историческом измерении, от древнерусской эпохи до наших дней.

Для В. фон Гумбольдта внутренняя форма – это тот способ, которым народный дух осознает мир, его «идеальное осмысление» [16, с. 103]. Соответственно, внутренняя форма слова – осмысление мира отдельно взятым словом. А.А. Потебня понимал под внутренней формой слова его этимологическое значение, которое образуется двумя компонентами (см. ниже). Внутренняя форма соответствует символу, поскольку в структуре символа так же содержится двойное означаемое. Рассмотрим определение, данное А.А. Потебней. В нем означаемому 1 и означаемому 2 символа соответствуют «сознание» и «содержание мысли»: «Внутренняя форма слова есть отношение содержания мысли к сознанию; она показывает, как представляется человеку его собственная мысль» [17, с. 91]. Эту, может быть, не вполне ясную формулировку А.А. Потебня сопровождает примером, демонстрирующим символический характер внутренней формы: для слова *туча* внутренняя форма предстает в виде отношения «мысль о туче» – «один из признаков тучи» (а именно то, что она «вбирает воду») [*там же*]. А.А. Потебня особо подчеркивает, что внутренняя форма выражает образ предмета, однако из этого образа выделяется единственный

признак [17, с. 90, с. 117], представляющий собой «центр образа» [17, с. 125]. Следовательно, символическая структура в процессе создания и функционирования внутренней формы реализуется следующим образом: слово (означающее 1) – один из признаков предмета, центр образа (означаемое 1, означающее 2) – мысль о предмете, образ предмета (означаемое 2).

Рассмотрим несколько дополнительных примеров, внутреннюю форму существительных *чудо*, *трудъ*, *печаль*. Праславянское слово **čudo* произведено от глагола **čuti*, означавшего ‘слышать, чувствовать, наблюдать’ [18, с. 128]. Внутренняя форма существительного определяет следующую символическую структуру: **čudo* (означающее 1) – ‘яркий признак объекта, воспринятый органами чувств’ (означаемое 1, означающее 2) – ‘любой объект, обладающий яркими, нестандартными признаками’ (означаемое 2). Вторичное означаемое соответствует в данном примере лексическому значению праславянского **čudo*, реальность этого исходного значения подтверждается зафиксированными в современных славянских языках значениями ‘странности’, ‘диковина’, ‘удивление’ (последние два значения находятся в отношениях метонимии) [18, с. 128].

Указанная сверхбинарная структура выявляется и при анализе внутренних форм двух других слов. Слово *печаль* восходит к глаголу **pekti (sę)*, у которого в ранних письменных памятниках зафиксированы только эмотивные значения ‘заботиться’, ‘скорбеть’ [19, с. 39]. Можно, однако, предполагать, что соответствующее физиологическое значение, ‘претерпевать жар’, было исходным для глагола. Исходя из этого, внутреннюю форму существительного следует определить следующим образом: **pečalь* (означающее 1) – ‘ощущение жара’ (означаемое 1, означающее 2) – ‘негативное переживание, связанное с высокой температурой’ (означаемое 2). Действие символического механизма в данном случае обеспечивается тем, что всякое эмоциональное состояние в праславянскую эпоху не существовало еще как самостоятельный феномен, а воспринималось посредством указывающих на него физиологических ощущений: боли, жара, тяжести, – которые являлись своеобразным симптомом, знаком душевного переживания.

Соответственно строилась и внутренняя форма слова *трудъ*. Индоевропейская глагольная база, на основе которой оформилось в праславянском языке данное существительное, выглядела как **treud* и обладала значением ‘мять, жать, давить, щемить’ [20, с. 266]. Ориентируясь на эту семантику и дальнейшую смысловую филиацию существительного, мы можем реконструировать символическую структуру, обусловленную внутренней формой: *трудъ* (означающее 1) – ‘ощущение тяжести’ (означаемое 1, означающее 2) – ‘неприятное состояние души и тела’ (означаемое 2). Во всех приведенных примерах внутренняя форма свидетельствует о языковом осмыслении действительности, позволяет судить об особенностях менталитета: восприятие человеком праславянской эпохи событий и явлений жизни предельно конкретно, основано на внимании к телесным ощущениям.

Внутренняя форма представлялась А.А. Потебне универсальным способом действием человеческого духа, он обнаруживал ее не только в слове и языке, но и в художественном произведении. Современные исследователи отмечают, что подобное широкое понимание внутренней формы поддерживается установленными закономерностями психической деятельности человека: всякое новое восприятие всегда обусловлено предшествующим эмоциональным, мировоззренческим, социальным опытом человека, его познаниями и сложившейся оценкой действительности. Таким образом, анализ внутренней формы неизбежно соотносит слово со всей областью культуры и учитывает основные функции символа, присутствующего «во всех доминионах культуры» [21, с. 13].

Традиция анализировать символ в культурологическом контексте была заложена в XX веке Э. Кассирером, который исследовал символические формы в различных сферах духовной деятельности: в языке, в мифологическом сознании, в процессах познания. Рассматривая символ как «смысловой знак», Э. Кассирер понимал его как финальную стадию развития форм языка [22, с. 247] и находил соответствие языковому процессу развития символов в области мифологических образов [22, с. 248]. Таким образом, символические свойства слова вводят его в пространство культуры и мифологии. В современных исследованиях языка принято говорить в этом случае о формирующейся с помощью слов языковой картине мира. Разумеется, слово при этом по-прежнему должно пониматься как символ, т. е. как образный знак, в структуре которого образ позволяет выразить дополнительное, не всегда ясное содержание.

Рассмотрим подробнее культурную значимость внутренней формы на примере этимологии слова *город*. Символический семиозис в этом случае может быть представлен в виде следующих связей: (1) **ge/ord* + **ŏ* <=> (2) ‘окружить, огородить’ + ‘результат процесса’ <=> (3) ‘огражденное, защищенное место’. Внутренняя форма (2), как и в предыдущих примерах, является означаемым для фонетической формы (1), но одновременно и означающим лексического значения (3). Именно в отношении (2) <=> (3) мы обнаруживаем образные качества символа, его мотивированность: неопределенное множество объектов (3) может быть представлено через признак замкнутой изгороди (2), и такой признак реально присутствует в каждом из называемых объектов. Реализуемый символ указывает на неопределенное, не вполне очерченное содержание (3), что, в частности, отличает символ от понятия. Неопределенность, синкретичность первоначального значения слова **gordъ* подтверждается данными славянских языков и диалектов, в которых этим словом (претерпевшим известные фонетические изменения) обозначается широкий спектр разнообразных объектов: городская стена, сад, крепость, замок, изгородь, хлев, конюшня, сарай, гумно, рынок, усадьба [23, с. 37-38]. Номинация каждого из объектов, указанных синкретичным лексическим значением, осуществлялась в праславянском языке через указание на существенный признак, соответствовавший внутренней форме, ‘замкнутая ограда’. Спектр называемых объектов, связанных с хозяйственной и социаль-

но-общественной инфраструктурой, оказывался достаточно широким, и потому в языковой картине мире формировалась важнейшая категория: противопоставление безопасного места дикой, необжитой территории.

В наступившую впоследствии письменную эпоху значение лексемы *градъ, городъ, gród* уточнялось в отдельных славянских языках, в результате чего в каждом из языков формировалось отдельное понятие и первоначальный признак внутренней формы включался в его содержание. В частности, в древнерусском языке были конкретизированы характеристики ограды через закрепление слова в определенных сочетаниях: *городъ с колиемъ, город камянъ, городъ каменыи, городъ твьрдъ, городъ земляной.* В результате этих и других уточнений оформилось древнерусское понятие *городъ*, включавшее различительные признаки ‘ограда’, ‘прочность, укрепленность’, а также ‘жители’, ‘центр власти’. На этом этапе слово как символ претерпело ослабление своих образных качеств и усиление знаковых свойств, что принято определять как затухание внутренней формы. Утрата образности – это утрата отношения (2) <=> (3), при которой элемент (2) совпадает с элементом (3). После такого совпадения может происходить полная утрата внутренней формы, что характерно для современного существительного *город*, у которого первоначальный признак внутренней формы не входит в минимальный набор необходимых признаков понятия. Соответственно, слово *город* с утратой внутренней формы перестает быть символом. Однако прекращается ли на этом семантическая эволюция слова?

Как мы показали выше, слово в своей символической функции становится важнейшим элементом культуры. С другой стороны, мы признаем, что символические структуры проявляются не только в языке, но на всех уровнях культуры. В частности, народные обычаи представляют символическую структуру следующим образом: ритуал или обряд (означающее 1) – первоначальный, мифологический смысл ритуала или обряда (означаемое 1, означающее 2) – обыденная, современная функция ритуала или обряда (означаемое 2). Именно поиском мифологического смысла обрядов, скрытого за современной функцией, занимался в своей известной работе А.Н. Афанасьев. В частности, при изучении обычаев, связанных с огнем, исследователь отмечал, что уверенность в лечебной силе огня восходит к архаичным похоронным ритуалам, в ходе которых проявлялось «очистительное значение огня, прогоняющего темную, демоническую силу и вместе с нею все греховное, нравственно нечистое» [24, с. 8]. По древнейшим мифологическим представлениям, огонь, разжигаемый при погребальном обряде, очищал умершего и предуготовлял его к жизни в лучшем мире. Структура символа может быть представлена в данном случае в следующем виде: [огонь] <=> ‘очищение умершего’ <=> ‘уничтожение всякого зла, в том числе болезни’.

Существительное *огонь* могло использоваться в заговоре, в тексте, сопровождающем лечебный обряд очищения огнем, ср.: «Ой, гульк вода! бусь вода! креши огню, жени биду!» [24, с. 8]. В результате слово *огонь* встраивалось в символическую цепочку, которая в приведенном примере, в соответствии с описанием А.Н. Афанасьева, принимает следующий вид: *огонь* (1) <=> [горячие угли] (2) <=> {‘молния Перуна’ (3) <=>} ‘средство от головной боли’ (4). В предложенной схеме денотат (2) является значением лексической единицы (1), но сам, в свою очередь, оказывается означающим для актуального содержания ‘лечение болезни’ (4), при том что внутренняя форма обряда (3), по-видимому, в значительной или в полной степени утрачена. Символ в данном примере проявляется в протяженной семантической цепочке.

Каким же образом подобное культурно значимое употребление слова отражалось на его семантической структуре? В случае с существительным *огонь* его лексическое значение не изменилось, а указание на лечебное средство осталось на уровне скрытой культурной коннотации. Однако в других случаях, когда слово участвовало в общекультурном символическом семиозисе, мог осуществляться перенос лексического значения.

Существительное *чудо* активно употреблялось в средневековых контекстах, связанных с описанием чудес, в житийных текстах и в специальных повестях о чудесах. В древнерусской культуре чудо – это необычайное, удивительное событие, в котором проявляется воля Господа, т. е. сама ситуация, обозначаемая словом *чудо*, выступала как означающее культурного знака, а весь семиозис строился по следующей схеме: *чудо* (1) <=> ‘яркое впечатление’ (2) <=> ‘объект удивления’ (3) <=> ‘проявление сверхъестественных сил – воли Господа’ (4). Внутренняя форма слова (2), возможно, еще осознавалась в древнерусском языке, однако впоследствии утратилась в связи с выходом из употребления производящего глагола *чути.* Значения (2) и (3) должны рассматриваться как прямое и переносное лексические значения, причем переход (2) => (3) был обусловлен функционированием культурного знака. Оба значения фиксируются в древнерусских памятниках [25, с. 1547], и оба они сохранились в современном русском литературном языке: «нечто небывалое, необычное, удивительное» – «сверхъестественное, необъяснимое явление, вызванное волшебством» [26, с. 1170]. Нужно уточнить, что символический семиозис слова *чудо* осуществляется только при условии, что сохраняется образность, которая связывает указанные два значения. То есть символ проявлялся в тех случаях, когда диво осознается как действие Бога, а деяние Бога – удивляет. Такая связь обязательна в религиозных контекстах, а в современной светской культуре древнерусский символ распадается на два автономных понятия, ‘нечто сверхъестественное’ и ‘нечто, удивляющее красотой’.

Перенос лексического значения мог происходить и в другом направлении, с означаемого культурного знака на его означающее, так произошло с существительным *держава.* В обряде венчания на царство в России с конца XVI века и в завершенной форме с начала XVII века в качестве символа власти использовались скипетр и держава («державное яблоко»). Реализуемая символическая структура выглядела при этом следующим образом: [предмет в форме шара с

крестом] <=> ‘вселенная’ <=> ‘власть над государством’. При употреблении в соответствующих контекстах слово включалось в семиотическую цепочку: *держава* (1) <=> [предмет в форме шара с крестом] (2) <=> ‘вселенная’ (3) <=> ‘власть над государством’ (4). Степень осознанности в то время внутренней формы (3) культурного знака нам неизвестна: на функционирование самого слова соответствующий признак никак не повлиял. Праславянское по происхождению существительное *держава* выражало значение ‘государственная власть’ (4) уже в ранних древнерусских памятниках [27, с. 222]. А вот указание на соответствующий предмет (2) оформилось именно в результате распространения новых элементов церемонии восшествия на престол – новое значение надежно фиксируется лишь в памятнике конца XVII века [там же].

Приведенные нами примеры показывают, что понятие символического семиозиса применимо не только к анализу внутренней формы, но и к анализу всего исторического развития слова. При семантической филиации лексемы (путем метафоры, метонимии, функционального переноса) старое значение становится мотивирующей базой для нового, в результате чего восстанавливается сверхбинарная структура символа, утраченная при затухании внутренней формы. Например, (1) *городъ* <=> (2) ‘защищенное место’ <=> (3) ‘население защищенного места’. Конечно, в реальном древнерусском контексте весьма часто (по распространенному мнению, в большинстве случаев) реализуются бинарные связи (1) <=> (2) или (1) <=> (3). В этих случаях принято говорить о прямом или переносном значении: при утрате опосредующего элемента слово, как и при утрате внутренней формы, сдвигается от символа к конвенциональному знаку.

Следовательно, интересующая нас проблема реализации символической функции слова сводится к вопросу о разграничении совмещенного, диффузного, символического значения (‘объект удивления + сверхъестественное явление’, ‘место’ + ‘население’) и значения прямого или чисто переносного, на основе которых формируется то или иное понятие (‘населенный пункт’, ‘население’, ‘диковина’, ‘сверхъестественное явление’). Наша гипотеза состоит в том, что символическое по функции, диффузное совмещение лексических значений возможно при взаимодействии между вербальным и другими культурными языками в процессе семантического развития слова. Предстоит, однако, большая работа для того чтобы определить, какие именно факторы способствуют реализации символических значений слова в системе язык – культура. По меньшей мере, в некоторых случаях символическая филиация слова может быть обусловлена воздействием знаков культуры на естественный язык. В московско-тартуской семиотической школе принято говорить в таких случаях о взаимодействии культурных кодов, однако мы позволим себе вслед за М.М. Бахтиным высказать некоторые сомнения в продуктивности термина «код». М.М. Бахтин отмечал, что идея кода «приводит к выпадению самого важного диалогического момента», т. е. препятствует взаимной проницаемости смыслов, так как код предполагает «какую-то готовость содержания и осуществленность выбора между данными кодами» [28, с. 339]. Знаковым средствам, составляющим код, соответствуют «фиксированные» значения [3, с. 34], поэтому затруднительно говорить о кодовых системах по отношению к символам, которые предполагают образную двуплановость, подвижность, неопределенность смысла.

Итак, мы рассматриваем слово как образный знак, который восстанавливает свою символичность в результате взаимодействия языков культуры: слово включается в символическую цепочку, когда начинает указывать на существующий культурный знак и, следовательно, становится знаком знака, т. е. символом. Именно эта особенность позволяет символическому качеству вторично – после внутренней формы – являться в слове. В ситуации, когда слово указывает на культурный знак, семиозис становится сверхбинарным. Подробнее такое восстановление символичности было нами рассмотрено по отношению к слову *чудо* в специальной работе [29]. Покажем далее некоторые полученные выводы.

В качестве означающего культурного знака в ситуации чуда могут выступать конкретные предметы и объекты (священная икона; монастырь, основанный чудотворцем) или события определенного рода (прежде всего, исцеление болезни, обычно слепоты). Слово, употребленное в описании чуда, становится знаком знака: (1) *чудо* <=> (2) объект или событие, означающее культурного знака, например, исцеление <=> (3) выражаемый культурным знаком смысл, например, ‘Божья воля, милость Господа’. Приведенная схема наглядно демонстрирует связь языка и культуры, а также возможности исследования смыслов культуры посредством описания лексических единиц. Отношение (1) <=> (2) представляет собой лексическое значение, традиционный предмет лингвистического изучения. Отношение (2) <=> (3) следует считать предметом изучения культурологической науки, а также, в тех случаях, когда соответствующий денотат предстает в виде образа в художественном произведении, – предметом литературоведческой науки. Вместе с тем в рассмотренном семиозисе присутствует также отношение (1) <=> (2) <=> (3): обязательным условием совершения чуда во всех случаях является молитва, явленная вера, т. е. осуществление связи с Господом – исцеление невозможно без Божьей воли, и слово *чудо*, указывая на необычайное событие, обязательно указывает на милость Господа.

Традиционная лексикология, придерживаясь узкой семантики, рассматривала в качестве лексического означаемого лишь непосредственный объект, названный словом, но не учитывала, что такой объект материальной реальности может выполнять знаковую культурную функцию. Поэтому указанное отношение (1) <=> (2) <=> (3) не было объектом лексикологического изучения: рассматривались значения прямые и переносные, но не символические. Диффузные, символические значения могут быть объектом исследования прежде всего исторической лексикологии, изучающей семантическую эволюцию лексем.

Слово в диахроническом аспекте представляет собой результат баланса между образными и знаковыми свойствами; символ обнаруживается в лексической семантике при анализе внутренней формы и культурных знаков. При символическом семиозисе происходит совмещение в пределах значения слова непосредственных материально-чувственных образов и выражаемого ими культурного содержания. При дальнейшем лексико-семантическом развитии освоенное содержание становится означающим для нового смысла, в результате формируется многочленная цепочка семиотических связей, напр.: *город* (1) <=> ‘нечто окруженное’ (2) <=> ‘крепостные стены’ (3) <=> ‘место для безопасной жизни, поселение’ (4) <=> ‘шум, суета’ (5). В подобную цепочку последовательно включаются внутренняя форма слова (2), древнерусское (архаичное) лексическое значение (3), современное актуальное значение (4), современные культурные ассоциации (5).

Постоянное удлинение подобной цепочки объясняется происходящим в истории слова ослаблением образности и усилением знаковости – мотивированность утрачивается, возникает потребность в непрерывном возобновлении образности и мотивированности, которые возрождаются при появлении новых означаемых. Следовательно, каждое прежнее означаемое может становиться означающим для нового означаемого, и так возникает глубокая смысловая перспектива, потенциально бесконечная цепочка значений, «все более абстрактных по мере удаления от исходного значения» [9, с. 133]. Таким образом, слово – это знак-символ, постоянно утрачивающий и обновляющий свою символичность. Утрата символичности превращает слово в конвенциональный знак, и знаковые свойства слова хорошо изучены. Символические же качества слова предстоит еще выявлять и анализировать на конкретном материале.

СПИСОК ЛИТЕРАТУРЫ

1. Тодоров Ц. Теории символов / пер. с фр. Б. Нарумова. – М.: Дом интеллектуал. кн., 1999. – 384 с.
2. Лотман Ю. М. Семиосфера . – СПб.: Искусство-СПб, 2001. – С. 12–149.
3. Чертов Л.Ф. Знаковость: опыт теоретического синтеза идей о знаковом способе информационной связи. – СПб.: Изд-во Санкт-Петербургского ун-та, 1993. – 378 с.
4. Флоренский П.А. У водоразделов мысли. Ч. 1. – М.: Правда, 1990. – 446 с.
5. Соссюр Ф. Заметки по общей лингвистике / пер. с фр. общ. ред. Н.А. Слюсаревой. – М.: Прогресс, 1990. – 280 с.
6. Соссюр Ф. Курс общей лингвистики. – Екатеринбург: Изд-во Урал. ун-та, 1999. – 425 с.
7. Степанов Ю.С. В мире семиотики // Семиотика: Антология / сост. и общ. ред. Ю. С. Степанов – 2-е изд., перераб. и доп. – М.: Академический проект; Екатеринбург: Деловая книга, 2001. – С. 5–42.
8. Соломоник А. Семиотика и лингвистика. – М.: Молодая гвардия, 1995. – 345 с.
9. Шелестюк Е.В. О лингвистическом исследовании символа (обзор литературы) // Вопросы языкознания. – 1997. – № 4. – С. 125–142.
10. Иванов Н.В. Символическая функция языка в аспектах семиогенеза и семиозиса: автореферат дис. ... доктора филологических наук. – М.: Воен. ун-т, 2002. – 34 с.
11. Якобсон Р. В поисках сущности языка // Семиотика: Антология / сост. и общ. ред. Ю. С. Степанов. – 2-е изд., перераб. и доп. – М.: Академический проект; Екатеринбург: Деловая книга, 2001. – С. 111–129.
12. Кибрик А.Е. Очерки по общим и прикладным вопросам языкознания: универс., типовое и специфич. в яз. – Изд. четвертое, стер. – М.: URSS, 2005. – 332 с.
13. Колесов В.В. Русская ментальность в языке и тексте. – СПб.: Петербургское востоковедение, 2007. – 619 с.
14. Рикёр П. Конфликт интерпретаций: очерки о герменевтике // Ин-т философии Российской академии наук / пер. с фр., вступ. ст. и коммент. И. С. Вдовиной. – М.: Академический проект, 2008. – 695 с.
15. Юнг К.Г. Архетип и символ. – М.: Ренессанс, 1991. – 297 с.
16. Гумбольдт В. фон. Избранные труды по языкознанию / пер. с нем. яз. под ред. и с предисл. Г.В. Рамишвили. – 2-е изд. – М.: Прогресс, 2000. – 396 с.
17. Потебня А.А. Мысли и язык. – М.: Лабиринт, 1999. – 268 с.
18. Этимологический словарь славянских языков. Вып. 4. – М.: Наука, 1977. – 232 с.
19. Словарь русского языка XI – XVII вв. Вып. 13. – М.: Наука, 1987. – 320 с.
20. Черных П.Я. Историко-этимологический словарь современного русского языка: в. 2-х т. – Т. 2. – М.: Русский язык, 1999. – 560 с.
21. Свасьян К.А. Проблема символа в современной философии: критика и анализ. – 2-е изд. – М.: Альма Матер: Академический проект, 2010. – 223 с.
22. Кассирер Э. Философия символических форм / пер. с нем. С.А. Ромашко. – Том 2. Мифологическое мышление. – М.; СПб.: Университетская книга, 2001. – 280 с.
23. Этимологический словарь славянских языков. Вып. 7. – М.: Наука, 1980. – 224 с.
24. Афанасьев А.Н. Поэтические воззрения славян на природу Опыт сравнит. изучения славян. преданий и верований в связи с миф. сказаниями других родств. народов. В 3 т. – Т. 2 . – М.: Соврем. писатель, 1995.– 396 с.
25. Срезневский И.И. Материалы для словаря древнерусского языка: в 3-х т. – Т. 3. – СПб., 1912.– 1684 с.

26. Словарь современного русского литературного языка: в 17-ти томах. – Т. 17. – М.-Л.: Наука, 1965. – 2126 с.
27. Словарь русского языка XI – XVII вв. Вып. 4. – М.: Наука, 1977. – 404 с.
28. Бахтин М.М. Эстетика словесного творчества / примеч. С. С. Аверинцева, С. Г. Бочарова. – 2-е изд. – М.: Искусство, 1986. – 444 с.
29. Алексеев А.В. Парадигматический и синтагматический аспекты лексикологического описания лексики переходного времени. Филологические традиции в современном литературном и лингвистическом образовании: сб. науч. ст. Вып. 10. В 3-х т. – Т. 2. – М.: МГПИ, 2011. – С. 111-115.

Материал поступил в редакцию 02.05.13.

Сведения об авторе

АЛЕКСЕЕВ Александр Валерьевич - кандидат филологических наук, доцент кафедры русского языка и общего языкознания Московского городского педагогического университета.

e-mail: alsalva@narod.ru

УВАЖАЕМЫЕ ЧИТАТЕЛИ!

ЦЕНТР НАУЧНО-ИНФОРМАЦИОННОГО ОБСЛУЖИВАНИЯ ВИНИТИ РАН

ПРЕДОСТАВЛЯЕТ КОПИИ ПЕРВОИСТОЧНИКОВ

ВИНИТИ РАН осуществляет обслуживание копиями первоисточников, хранящихся в фонде научно-технической литературы ВИНИТИ, в фондах других библиотек, а также в доступных ВИНИТИ электронных ресурсах.

Фонд научно-технической литературы ВИНИТИ включает более 2 млн изданий по точным, естественным и техническим наукам, в т.ч.:

– отечественные и иностранные периодические и продолжающиеся издания – с 1987 г.;

– отечественные книги – с 1987 г.;

– иностранные книги – с 1991 г.;

– рукописи, депонированные в ВИНИТИ, – с 1962 г.

Заказы на бумажные или электронные копии первоисточников принимает Центр научно-информационного обслуживания (ЦНИО) ВИНИТИ. ЦНИО ВИНИТИ обслуживает коллективных (организации и учреждения) и индивидуальных пользователей.

Формы обслуживания:

– абонементная (на основе договоров и предоплаты);

– разовые заказы (с предоплатой заказа по счету);

– индивидуальная форма обслуживания в читальном зале ЦНИО ВИНИТИ.

На сайте ВИНИТИ (http://www.viniti.ru) представлен полный Электронный каталог научно-технической литературы (http://catalog.viniti.ru), зарегистрированной в ВИНИТИ с 1994 г. Доступ для просмотра и поиска по Каталогу свободный. Постоянные абоненты ЦНИО ВИНИТИ, имеющие логин и пароль для работы с Каталогом, могут делать заказ копий непосредственно через Каталог.

Услуги по изготовлению копий первоисточников из фондов других библиотек предоставляются только постоянным абонентам. Место хранения первоисточников указывается в Электронном каталоге.

За подробной информацией обращаться по адресу:

125190, Россия, Москва, ул. Усиевича, 20, ВИНИТИ РАН. ЦНИО

Телефоны: 8 (499)155-42-43, 155-42-09, 152-54-59

Факс: 8 (499) 943-00-60

E-mail: cnio@viniti.ru; **URL:** http://www.viniti.ru

РОССИЙСКАЯ АКАДЕМИЯ НАУК

Федеральное государственное бюджетное учреждение науки
ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК

предлагает научным работникам, аспирантам и другим специалистам в области естественных, точных и технических наук, желающим быстро и эффективно опубликовать результаты своей научной и научно-производственной деятельности, использовать способ публикации своих работ через *систему депонирования.*

«Депонирование (передача на хранение) – особый метод публикации научных работ (отдельных статей, обзоров, монографий, сборников научных трудов, материалов научных мероприятий – конференций, симпозиумов, съездов, семинаров) узкоспециального профиля, разрешенных в установленном порядке к открытому опубликованию, которые нецелесообразно издавать полиграфическим способом печати, а также работ широкого профиля, срочная информация о которых необходима для утверждения их приоритета. Депонирование предусматривает прием, учет, регистрацию, хранение научных работ и обязательное размещение информации о них в специальных информационных изданиях».

Подготовка и передача на депонирование научных работ происходит в соответствии с «Инструкцией о порядке депонирования научных работ по естественным, техническим, социальным и гуманитарным наукам» (М., 2003).

Результатом депонирования является публикация информации о депонированных научных работах в информационных изданиях ВИНИТИ РАН – Реферативном журнале и аннотированном библиографическом указателе «Депонированные научные работы».

В соответствии с “Положением о порядке присуждения ученых степеней”, утвержденным Постановлением Правительства Российской Федерации от 30.01.2002 № 74 (в ред. Постановлений Правительства РФ от 20.04.2006 № 227, от 02.06.2008 № 424, от 20.06.2011 № 475), научные работы, депонированные в организациях государственной системы научно-технической информации, признаны публикациями, учитывающимися при защите кандидатских и докторских диссертаций.

Подать научную работу на депонирование можно обратившись в Отдел депонирования ВИНИТИ РАН по адресу:

125190, Москва, ул. Усиевича, 20.
ВИНИТИ РАН, Отдел депонирования научных работ.
Тел.: 8 (499) 155-43-28, Факс: 8 (499) 943-00-60.
e-mail: dep@viniti.ru

С инструкцией о порядке депонирования можно ознакомиться на сайте ВИНИТИ РАН: **http://www.viniti.ru**