Создание, поддержка и развитие модели интерпретации смыслов

В. Н. Добрынин[1,a], И. А. Филозова[2,b]

[1] Государственный университет «Дубна»,
141980, Россия, Московская обл., г. Дубна, ул. Университетская, 19

[2] Объединенный институт ядерных исследований,
141980, Московская обл., г. Дубна, ул. Жолио-Кюри, 6

E-mail: [a] arbatsolo@yandex.ru, [b] fia@jinr.ru

В мире наблюдается постоянный и устойчивый рост количества научных рецензируемых журналов и публикуемых в них статей, а также накопление контента в специализированных информационных фондах. Хотя значительная часть такого рода ресурсов находится в электронных фондах, тщательно изучать такие информационные массивы для ученых и исследователей традиционными способами становится все труднее. Но смысловой поиск – это необходимый этап, предшествующий генерации нового знания в научной среде, и порядка 60% времени ученый тратит именно на поиск научной информации. Научные статьи представляют собой тексты на естественном языке, отвечающие определенным требованиям к структуре и содержанию: однозначность, логика от посыла к следствию, явная целевая установка статьи, ясность и точность. Несмотря на шаблон, в них присутствует неопределенность, связанная с неоднозначностью интерпретации читателем.

В работе излагается технология извлечения смысла из научных текстов на основе модели интерпретации (донесения до определенной аудитории) смысла, состоящей из следующих компонентов: выделение смысла из научной статьи на основе построения выжимки, словаря, семантической модели (слова и их взаимосвязи); формирование логико-семантической модели (сети); вопросно-ответный параметрический навигатор. Такая модель используется для структурирования информационных фондов на основе каталожной службы, представляющей собой множество логико-семантической сетей. Эффектом применения такого подхода является сокращение времени на изучение фонда за счет повышения уровня понимания данной статьи.

Ключевые слова: электронный фонд, научный стиль речи, интерпретация, логико-семантическая сеть, вопросно-ответный навигатор

1. Введение

По результатам некоторых исследований к началу 2014 г. в мире было выпущено 34 274 рецензируемых журнала [Домнина, Хачко, 2015]. База данных Journal Citation Reports медиакомпании Thomson Reuter's (США) включает более 12 700 журналов, публикующих ежегодно 1 200 000 статей. База данных Scopus (Нидерланды, Elsevier) охватывает более 21 000 журналов. По данным реестра репозиториев открытого доступа (http://roar.eprints.org/) в мире на данный момент насчитывается 4359 институциональных архивов с суммарным числом записей порядка 42 млн. Поскольку генерация нового знания в некоторой предметной области неизбежно предшествует поиску необходимой информации, специалист-исследователь, решая свои профессиональные задачи, сталкивается с проблемой семантического поиска в больших информационных массивах. И осуществить такой поиск с требуемым уровнем качества за приемлемое время довольно непросто. Зачастую результаты поиска не удовлетворяют в полной мере информационную потребность пользователя. Одной из причин этого явления является неточное и/или неполное представление пользователя о том, что должно являться результатом поиска. Таким образом, представляется актуальной задача повышения эффективности исследования научных электронных информационных фондов читателем за счет снижения неопределенности интерпретации им научного текста.

2. Модель интерпретации смыслов, извлеченных из научного текста

2.1. Взаимосвязь данных, информации, смысла и знаний

Общеизвестно, что понятия данные, информация, смыслы и знания не являются синонимами, и не существует однозначных определений этих терминов. В данном случае, будем придерживаться следующих трактовок. *Данные* − это совокупность сведений, зафиксированных на определенном носителе в форме, пригодной для постоянного хранения, передачи и обработки. *Информация* − это результат преобразования и анализа данных. З*нание* (в узком смысле) — обладание проверенной информацией, позволяющей решать поставленную задачу. Для ее решения данные обрабатываются на основании имеющихся знаний, затем происходит анализ полученной информации с учетом имеющихся знаний. Далее на основании анализа предлагаются все допустимые решения для последующего наилучшего в некотором смысле решения. Результаты решения пополняют знания.

Когда источником информации является текст, возникает задача извлечения смысла, т.е. его (текста) понимания, которая может оказаться трудной в силу смысловой загруженности текста, употребления специальных терминов и/или большого объема. В данном случае смысл — это главная мысль, идея высказываний (например, научной статьи).

Понимание рассматривается в герменевтике (деятельность человека или коллектива при понимании или интерпретации текста или объекта, который может трактоваться как текст) как одно из инобытий рефлексии, а высказанная рефлексия есть интерпретация. Специалисты по герменевтике выделяют более сотни техник понимания, в т.ч. интерпретационного типа [Богин, 2001]. ИНТЕРПРЕТАЦИЯ (лат. interpretatio): в широком смысле – истолкование, объяснение, перевод на более понятный язык; в специальном смысле – построение моделей для абстрактных систем (исчислений) логики и математики [Большой энциклопедический словарь, 2012]; а также процесс толкования и разъяснения смысла чего-то неясного или сложного для понимания кому-либо. К техникам понимания относят, например, растягивание смыслов − их категоризация, переход от собственно смыслов к метасмыслам; выход к пониманию как осознанному усмотрению и/или построению смысла; восстановление смысла по значению; самоопределение в мире усмотренных смыслов; движение понимание → интерпретация → дальнейшее понимание (и далее); оценка собственного понимания в связи с определением средств текста, обеспечивших пробуждение рефлексии. Техники могут сочетаться самым разнообразным способом,

образуя сложнейшую схему. Использование той или иной техники понимания предполагает усилия понимающего субъекта, который должен либо дискурсивно построить для себя вопросы, либо вопросы уже поставлены перед ним, и он вынужден искать на них ответы [Богин, 2001].

Восприятие смыслов текста (в данном случае научной статьи) есть "постижение смысла того или иного явления, его места, его функции в системе целого" [Коршунов, Мантатов 1986]. Тогда смысл текста есть результат процесса его восприятия субъектом, т.е. создаваемая им "идеальная конструкция", явно не содержащаяся в тексте. "С этой точки зрения текст можно рассматривать как программу, по которой можно построить некоторое число смыслов" [Шукуров, Нишанов 1983].

Таким образом, грань между информацией и знаниями нечеткая и зависит от воспринимающего субъекта.

2.2. Формулировка задачи

С другой стороны, данные — это формальные символические структуры, для которых может быть определён синтаксис, семантика и логические правила преобразования этих структур. Если этим структурам определить смыслы и систему правил соответствия, то возникают две фундаментальные проблемы:

1. Заданы данные. Как определить смысл этих данных? Т.е. задать интерпретацию данных как процесс, обеспечивающий получение смысла из данных.
2. Обратная задача. Задан смысл. Как по смыслу создать формальную систему (язык, т.е. синтаксис, семантику)?

В данном случае рассматривается первая задача. Способ решения состоит в построении технологии извлечения смысла из данных, т.е. модели интерпретации данных.

2.3. Структура модели

В качестве исходных данных выступают электронные информационные фонды научных текстов. Все множество текстов можно условно разделить на три категории:

1. Неструктурированные (тексты на естественном языке);
2. Слабо структурированные − тексты на естественном языке, но подчиняющиеся правилам (структурным, лексическим и т.д.);
3. Хорошо структурированные тексты (формальные тексты, в том числе и данные).

Научные тексты относятся ко второй категории. Им присущи такие признаки, как логичность изложения, смысловая точность (однозначность), обобщенность, объективность. Научный стиль имеет лексические, морфологические, синтаксические, структурные особенности, а также характерное использование средств выразительности. Таким образом, они подчиняются некоторому шаблону.

Отсюда следует структура модели, а именно: *Объект* (это может быть текст, речь, реальные или идеальные образы, изображения и т.д.) не ясен *Субъекту1* с точки зрения понимания, смысла и т.д.). Тогда *Субъект2* толкует *Объект* так, что *Субъекту*1 неясное становится доступным в понимании. Таким образом, для выявления пространства смыслов строится модель их интерпретации, направленная на понимание смысла, заложенного автором.

2.4. Семантическое структурирование электронных научных информационных фондов

Построение модели предлагается проводить на основе каталожной службы, ориентированной на пространство смыслов, суть которой состоит в следующем. Любая задача, научно-технический текст или образовательный ресурс могут быть представлены логической последовательностью вопросов и ответов, дополненной смысловым описанием (реакцией). Основой технологии является способ описания предметной области и информационных ресурсов мно-

жеством логико-семантических сетей (ЛСС) «вопрос-ответ-реакция» [Добрынин, Филозова, 2014]. Такая сеть — это упорядоченное множество вопросов, ответов и связей между ними, образующее целостную систему (рис.1).

Рис.1. Формальная структура и связь вопроса и ответа

Итак, научно-практическая область знаний включает предмет исследования и проблемное поле (перечень проблемных вопросов). Проблемные вопросы представлены иерархическим деревом по принципу «от общего к частному». Для некоторых вопросов уже существуют возможные альтернативные ответы и способы их реализаций (реакции). Реакции вопроса — это описание области предпосылки вопроса для осознания обстоятельств и причин возникновения вопроса и дальнейшего установления смыслового соответствия с областью ответа. Реакции ответа — это описание области ответа для осознания смысла вопроса и смысловой связи с ответом; смысловое пространство, имеющее связь с пространством вопроса, из которого следует ответ. Ответы могут порождать в свою очередь вопросы (рис.2.).

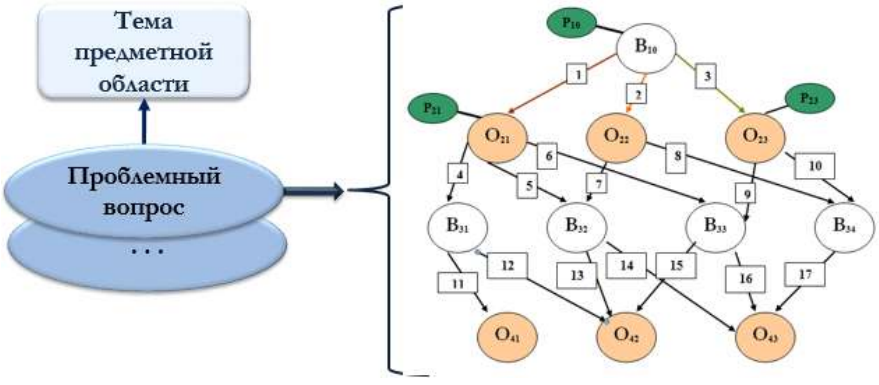

Рис. 2. Граф ЛСС, раскрывающий проблемный вопрос

Реакции играют важную роль в системе, а именно, очерчивают границы поля смыслов, потенциальных для извлечения. Таким образом, знания, накопленные в предметной области, могут быть представлены множеством ЛСС «вопрос-ответ-реакция», упорядоченных по предметным темам. Единицей ЛСС является логическая связка ВОПРОС-ОТВЕТ и связанные с ними

реакции. Поскольку ЛСС представляет собой направленный граф, это обеспечивает механизм навигации по выбранному маршруту. Построение множества ЛСС возможно на нескольких уровнях (слоях), охватывающих собственно информационный ресурс (научную публикацию), тему предметной области, предметную область в целом и понятийный слой – словари, тезаурусы, онтологии.

Движение вниз по сети углубляет знания, по горизонтали расширяет знания, вверх − обобщает знания. С точки зрения пользователя это означает изучение информационного ресурса в режиме вопрос-ответ.

3. Реализация подхода

Реализация подхода включает разработку следующего инструментария (ПО): 1) автоматизированного рабочего места (АРМ) аналитика для структурирования информационного фонда, предназначенного для создания и редактирования множества ЛСС; 2) вопросно-ответного навигатора, обеспечивающего движение по многоуровневой сети ЛСС. В основе разработки — методика анализа научных текстов, включающая набор специальных фильтров:

- Фильтр 1. Общая часть — анализ проблемы, ее история, обзор, актуальность.
- Фильтр 2. Авторские понятия — вводимые авторами новые термины, обще-употребляемые термины с авторской интерпретацией, сужающие семантику.
- Фильтр 3. Примеры и иллюстрации для пояснения сложных мест в тексте, сокращение размера текста при строгих ограничениях по объему.
- Фильтр 4. Идея автора — описание и раскрытие основной авторской идеи.

На основе полученной разметки эксперт производит формирование вопросов, ответов и реакций, а затем — связывание их в ЛСС (рис.3.).

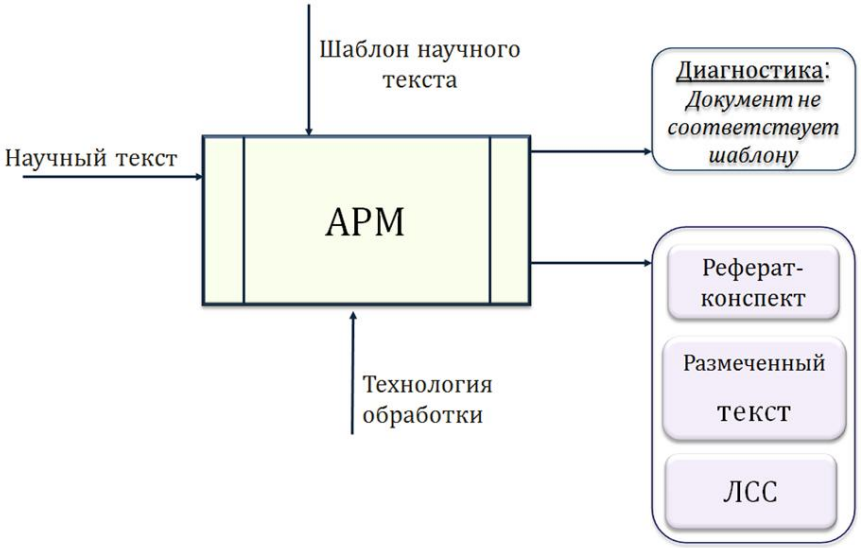

Рис.3. Функционирования АРМ аналитика

Аналитику фонда доступны следующие функции: анализ текста, создание ЛСС, редактирование ЛСС, просмотр ЛСС. На этапе анализа текста эксперт производит разметку текста на основе выше описанных фильтров. Результаты разметки сохраняются в базе данных. На этом материале формируются список вопросов с реакциями и список ответов с реакциями (рис.4.). Затем из блоков вопрос-ответ-реакция формируется ЛСС.

На данный момент реализованы прототипы АРМ аналитика и вопросно-ответный навигатор, апробированные на корпусе научных документов, размещённых на тестовом экземпляре сервиса «JINR Document Server» (репозитория Открытого Доступа статей, препринтов и других

материалов о научно-исследовательской деятельности в ОИЯИ), развернутого в облачной инфраструктуре ОИЯИ [Baranov et al., 2016].

Пользовательский интерфейс для просмотра информационного ресурса, каталогизированного на основе ЛСС, организован в виде набора информационных карточек вопросов (рис. 5).

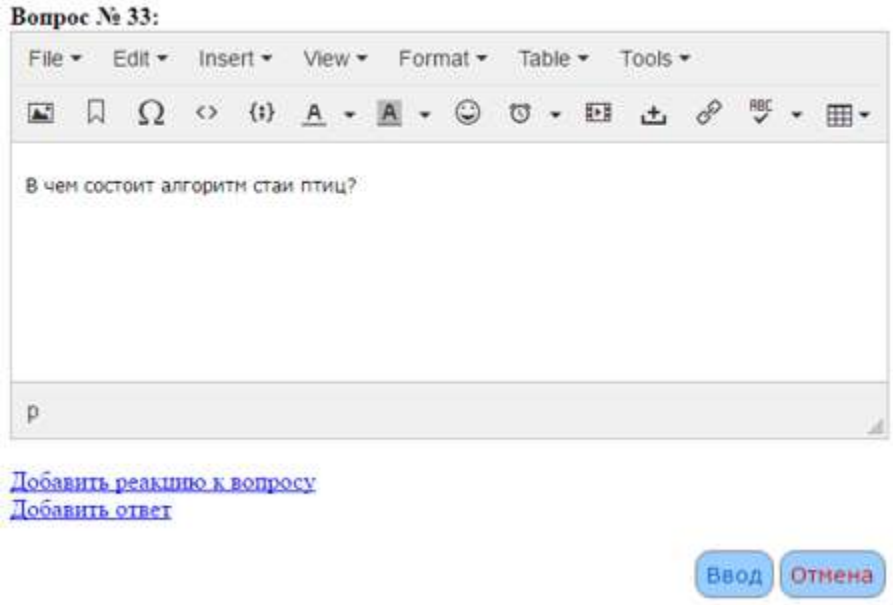

Рис.4. Формирование вопросов, ответов и реакций

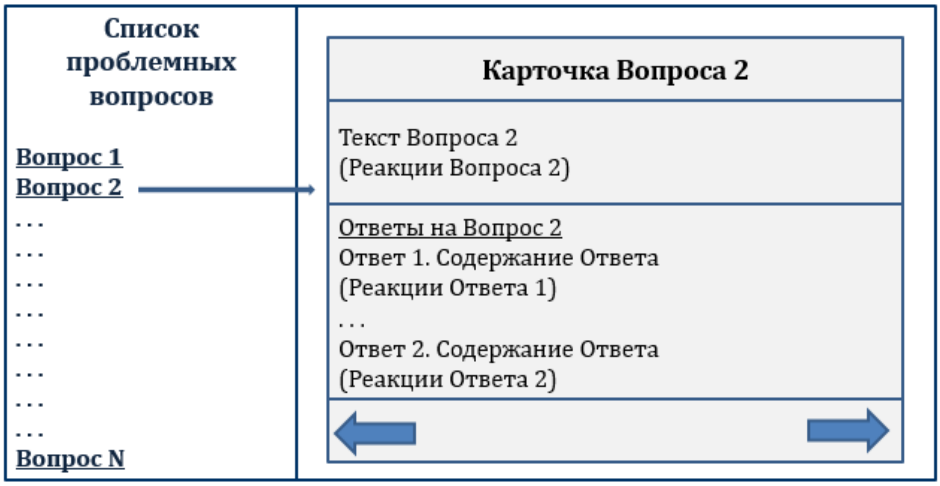

Рис.5. Общий вид интерфейса пользователя в режиме чтения

На рис.6. представлен фрагмент страницы с вопросно-ответным представлением, сформированным для издания *О.Д. Маслов и др. Извлечение урана из зольных отвалов от сжигания углей Монголии / О.Д. Маслов ; Ш. Цэрэнпил ; Н. Норов ; М.В. Густова // Дубна: ОИЯИ, 2010. - 7 с. [JINR-P12-2010-27]*

4. Заключение

Технология работает с классом научных текстов, для которых определены шаблон и технология извлечения явных и скрытых смыслов. Модель есть интерпретатор текста, цель которого выявить смысл научного текста и сделать его доступным для определённой аудитории.

Данный подход может применяться для структурирования образовательного контента для обеспечения самостоятельного эффективного изучения материала обучающимися; как компонент интеллектуальных информационных систем (консультативных, агентных, социальных и т.д.).

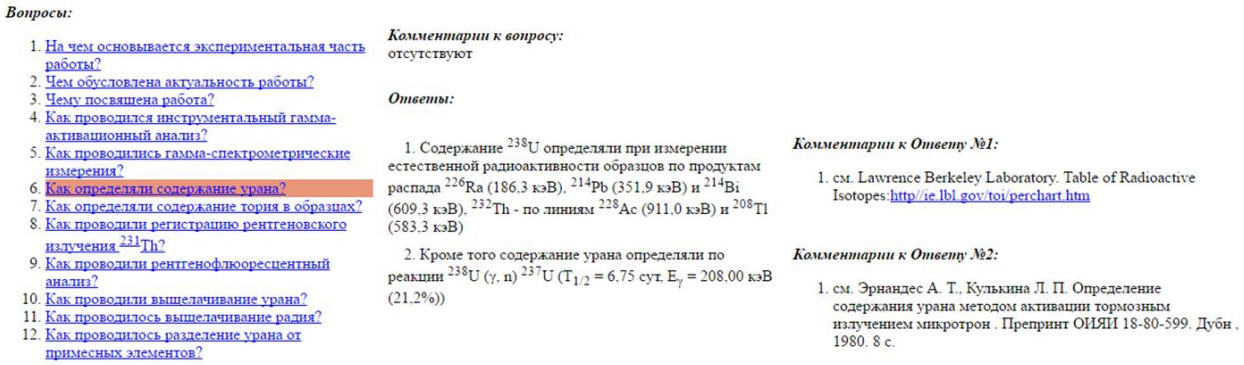

Рис.6. Фрагмент страницы для просмотра ресурса в режиме вопрос-ответ

Список литературы

Домнина Т. Н., Хачко О. А. Научные журналы: количество, темпы роста // Информационное обеспечение науки: новые технологии, Сб. науч. тр. / Каленов Н.Е., Цветкова В.А. (ред.). - М.: БЕН РАН, 2015. - с.83-96.

Domnina T.N, Khachko O. A. Nauchnye zhurnaly: kolichestvo, tempy rosta [Scientific journals: the amount, the growth rate] // Informatsionnoe obespechenie nauki: novye tekhnologii, Sb. nauch. tr. / Kalenov N.E., Tsvetkova V.A. (red.). – M.: BEN RAN, 2015. – c.83-96 (in Russian).

Большой энциклопедический словарь. 2012 [Электронный ресурс]: https://slovar.cc/enc/bolshoy/2087755.html

Богин Г.И. Обретение способности понимать: Введение в филологическую герменевтику.— Москва, 2001.

Bogin G.I. Obretenie sposobnosti ponimat': Vvedenie v filologicheskuyu germenevti-ku.— Moskva, 2001 (in Russian).

Коршунов А. М. Мантатов В. В. Гуманитарное знание и понимание // Философские науки, 1986, № 5.

Korshunov A. M. Mantatov V. V. Gumanitarnoe znanie i ponimanie // Filosofskie nauki, 1986, № 5 (in Russian).

Шукуров Э. Д., Нишанов В. К. Семасиогенезис и проблема понимания//Философско-методологические проблемы теории общения. - Фрунзе, 1983.

Shukurov E. D., Nishanov V. K. Semasiogenezis i problema ponimaniya//Filosofsko-metodologicheskie problemy teorii obshcheniya. - Frunze, 1983 (in Russian).

Добрынин В.Н., Филозова И.А. Семантический поиск в научных электронных библиотеках // Информатизация образования и науки № 2(22) / 2014. - с.111-127

Dobrynin V.N., Filozova I.A. Semanticheskii poisk v nauchnyh electronnyh bibliotekakh [Semantic search in the scientific digital libraries] // Informatizatsija obrazovanija i nauki № 2(22) / 2014. - c.111-127 (in Russian).

Baranov A.V. et al. JINR cloud infrastructure evolution / A.V. Baranov, N.A. Balashov, N.A. Kutovskiy, R.N. Semenov // Particles and Nuclei Letters, v.13, No. 5, pp.1046-1050, 2016

Creating, Supporting and Developing of Model of Meanings Interpretation

V. N. Dobrynin[1,a], I. A. Filozova[2,b]

[1] Dubna State University, Universitetskaya, 19, Dubna, Moscow region, 141980

[2] Joint Institute for Nuclear Research, Joliot-Curie, 6, Dubna, Moscow region, Russia, 141980

E-mail: [a] arbatsolo@yandex.ru, [b] fia@jinr.ru

We observe a continuous and steady growth in the number of peer-reviewed scientific journals and published articles in the world. Information content is accumulated in special funds. The large amount of these resources is presented in the digital funds. However the careful studying such information arrays for scientists and researchers is becoming more difficult by standard methods. But semantic search is a necessary step before the generation of a new knowledge in the scientific community, and about 60% of the time scientist spends for the search of the need information. Scientific articles are texts in natural language that correspond to certain requirements to the structure and content: uniqueness, the logic of the discourse to the effect, a clear mission of article, clarity and accuracy. Although the pattern such texts have uncertainty because of the ambiguity of the interpretation by the reader.

The paper describes the technology of the meaning extraction from scientific texts based on a model of interpretation (explanation for a specific audience) interpretation of the meaning, consisting of the following components: the extraction of the meaning of a scientific article based on the building the summary, vocabulary, semantic model (words and their relations); the formation of logical-semantic model (network); question-answer parametric navigator. This model is used to structure the information funds based on catalog service that is a set of logical-semantic networks. The usage effect of this approach is to reduce the time to study the fund by raising the level of understanding of this article.

Keywords: digital fund, scientific style of discourse, interpretation, logical-semantic network, question-answer navigator