

# ФОРМИРОВАНИЕ МОДЕЛИ ПРЕДМЕТНОЙ ОБЛАСТИ НА ОСНОВЕ ВТОРИЧНЫХ ДОКУМЕНТОВ В ЭЛЕКТРОННО-ПОИСКОВОЙ СРЕДЕ

Е.И. Моисеев, А.А. Муромский, Н.П. Тучкова

Работа посвящена проблеме представления предметной области «свертывание научной информации» в электронной информационной среде. Известно, что с начала 90-х гг. XX в. сеть ИНТЕРНЕТ и WEB-порталы по сути своей являются системами, реализующими технологию интеграции данных [1], поступающих из различных областей знаний, на основе иерархических и ассоциативных связей, и во многом заменяющими научному сообществу остальные средства коммуникации. Поэтому идея создания моделей знаний в этой среде составляет часть общего планирования дальнейшего развития сети ИНТЕРНЕТ. Представление предметных областей в ИНТЕРНЕТ предполагает междисциплинарные исследования, как самой предметной области и смежных с ней областей, так и средств реализации информационной модели. Создание концептуальных моделей разделов предметных областей и формирование на этой базе тезаурусов и онтологии предметных областей составляет предмет исследования авторского коллектива. Исследования направлены на решение задачи интеграции данных и научных результатов в сети ИНТЕРНЕТ на основе анализа научных публикаций и разработки информационной модели предметной области, включая дескрипторные словари, указатели ключевых слов, авторские и тематические указатели и предоставлении научному сообществу информационного ресурса для формирования запроса в сети.

Создание семантических моделей предметных областей и их приложений в виде разделов в информационной электронной среде способствует устранению дублирования понятий, формированию массивов эквивалентных русско-английских терминов предметных областей, а также служит систематизации знаний в электронной среде в целом.

Для построения связей используются парадигматические отношения, понятие тезаурус адресата и пр. [2]. В процессе работы формируется словарь терминов предметных областей и буквенно-цифровых кодов – идентификаторов, с помощью которых возможно внедрение контролируемой лексики в процессы ведения и поддержки баз данных.

Фактически каждый пользователь системы ИНТЕРНЕТ на основе фиксированного массива компонентов, предлагаемого в данной работе, может по-разному, индивидуально представить свои труды коллегам-пользователям в системе ИНТЕРНЕТ. Фиксированное множество компонентов при использовании их для представления разных областей знания в виде вторичных документов вносит так необходимые при поиске информации элементы стандартизации, о чем неоднократно писали разработчики ГРИД-технологии [3].

## Информация в научной организации и вторичные документы

Под информацией в научной организации прежде всего понимаются научные результаты коллектива, публикуемые в изданиях организаций, в типографиях институтов, или где-нибудь, но как труды организаций. Классическим примером такой научной организации является Вычислительный центр им. А.А.Дородницына (ВЦ РАН). Как известно, в ВЦ РАН публикуются следующие виды изданий: Сообщения ВЦ РАН, Монографии ВЦ РАН, Сборники ВЦ РАН. На протяжении ряда лет выпускался УКАЗАТЕЛЬ трудов ВЦ РАН [4], который является одним из вариантов вторичных документов (secondary documents) первоисточников – сообщений, монографий, статей сборников в ВЦ РАН [5]. Вторичные документы однозначно соответствуют первичным документам. Отметим, что качество вторичного документа зависит также от качества первоисточника, как научного документа, если создатель вторичного документа ограничивается только рамками данного первоисточника, т.е. не рассматривает первоисточник во всей данной теме, не внося соответствующее во вторичный документ в аспекте сравнения, значимости и т.д. Такая работа предполагает соответствующий уровень профессионализма у разработчика вторичного документа, и в этом случае качество вторичного документа зависит уже не только от первоисточника, но и от профессиональной подготовки его составителя.

Формирование массива вторичных документов рассматривается как решение задачи свертывания научной информации. Заметим также, что специфика самого процесса формирования массива вторичных документов такова, что чем больше этот массив, тем больше связей надо учитывать при его пополнении, но при этом в полученный информационный ресурс добавляется больше структурированной и систематизированной информации. Таким образом, инструмент вторичных документов составляет одну из основ для описания понятий и связей предметных областей. Например, используя компоненты УКАЗАТЕЛЯ ВЦ РАН, получаем представление о политематике ВЦ РАН.

Рассмотрим пример на основе сборника ВЦ РАН: Исследование операций (модели, системы, решения). Сб. статей. Отв. редактор доктор физ.-матем. наук А.П.Абрамов. М.: ВЦ РАН, 2006. 84 с. ISBN 5-201-09844-4.

Библ. описание	Исследование операций (модели, системы, решения). Сб. статей. Отв. редактор доктор физ.-матем. наук А.П.Абрамов. М.: ВЦ РАН, 2006. 84 с. ISBN 5-201-09844-4.
----------------	--

Аннотация	В сборник включены работы по применению оптимального управления к задачам химической кинетики, исследованию модельных оценок в диффузионной модели корреляционной адаптометрии, моделированию популяционных циклов с помощью индивидуально-ориентированной модели, идентификации консервативных по Вольтерра матриц использованию кластер-анализа в формировании научных коллективов, применению методов корреляционной адаптометрии для оценки эффективности лечения некоторых болезней, а также использованию ЭВМ в количественной экологии	
Рецензенты	Б.В. Архипов, А.П. Афанасьев	
Оглавление	.....	
гиперссылка из оглавления с. 28-33	Использование методики корреляционной адаптометрии для оценки эффективности лечения больных ожирением.	
	Гиперссылка	<b>Медицина. Охрана здоровья.</b> (Охрана здоровья № ...)
	<b>Тематический Указатель</b>	<b>Математические методы в медицине</b> (Медицина №...)
	Авторы	<u>В.Н. Разжевайкин,</u> <u>М.И. Шпитонков,</u> <u>А.В. Васильев,</u> <u>Г.Ю. Мальцев,</u> <u>Ю.В. Хрущева</u>
	Аннотация	Методика корреляционной адаптометрии была использована для исследования группы 70 больных в возрасте от 18 до 60 лет, страдающих ожирением 1-3 степени алиментарного генеза без сопутствующих заболеваний и с наличием осложнений функционального или органического характера. Состав тела пациентов оценивали методом биоимпедансного анализа с помощью анализатора «ABC-01 MEDACC» (РФ). Биохимические показатели крови определяли на анализаторе открытого типа Канелаб. Приведены результаты исследований и их анализ
<b>Ключевые слова.</b>	Больные ожирением, Многопараметрические исследования больных, Корреляционная адаптометрия, Терапия методом корреляционной адаптометрии, Корреляционный граф, Попарные коэффициенты корреляции, Вес корреляционного графа, G-критерий, Оценки связности физиологических параметров, Оценки здоровья населения, Индивидуальные программы питания	
Литература	.....	

Использование массива вторичных документов для авторского и тематического поиска

Приведенный выше пример носит типический характер. Здесь не играет роли тематика и специфика терминологии. На основе УКАЗАТЕЛЯ трудов ВЦ РАН составлен массив из 60 компонентов, который практически может быть использован для описания научных результатов в печатной и электронной информационно-поисковой среде по всей политематике ВЦРАН (вычислительная математика, численный анализ, вычислительная техника, теория, технология и применения вычислительных машин и систем, гидродинамика, механика газов, аэродинамика, декомпозиционные методы, исследование операций, газовая динамика, математическая физика, общая механика, оптимизация, распознавание и преобразование образов, экология, разработка нефтяных и газовых месторождений и др.).

Развитые методы представления научных публикаций могут использоваться индивидуально каждым пользователем ИНТЕРНЕТ. Множество компонентов может пополняться, так как является открытым множеством. Это позволяет реализовать процедуру свертывания научной информации в любой тематике. Для этого необходимо, используя средства HTML, разместить вторичный документ публикации на персональную страницу автора или организации, в которой опубликована работа с соблюдением всех юридических норм. Систематическое пополнение списков вторичных документов публикаций (авторских и/или в рамках организации) фактически приведет к формированию информационной модели предметных областей как авторских исследований, так и организации в целом. Данные будут доступны в известных поисковых системах, что будет способствовать росту цитирования.

Далее, для выявления высокоинформативных публикаций могут быть применены библиометрические исследования, результаты, которых могут быть также представлены авторам и использованы в работе.

Предложенные методы укладываются в рамках активно развивающейся тематики – формирование информационного пространства РАН. Предполагается, что результаты исследований будут способствовать уменьшению имеющегося разрыва между теорией семантического подхода в обработке информации и практикой формирования информационных моделей (таксономий, тезаурусов и онтологий) конкретных предметных областей, например [6,7]. Все полученные результаты предполагается обобщить в

информационном ресурсе, описывающем выбранные предметные области, для представления в интегрированной информационно-поисковой распределенной среде РАН с помощью средств RDF-схем языка OWL.

Работа выполнена при поддержке Российского Фонда Фундаментальных Исследований (проект 07-07-00082а).

#### ЛИТЕРАТУРА:

1. Berners-Lee, T., Hendler, J., and Lassila, O. The semantic Web // Scientific American Magazine. May 2001
2. Шрейдер Ю.А. Об одной модели семантической теории информации // Проблемы кибернетики. Вып. 13. М.: «Наука», 1965. С. 223-240
3. Технологии GRID. Том 1,2. М.: ИПМ им. М.В.Келдыша, 2006. 380 с. ISBN 5-98354-003-3.
4. УКАЗАТЕЛЬ-XXXX(XXXX=2000, 2001,2002,2003,2004,2005). Указатель трудов Вычислительного центра им. А.А.Дородницына Российской академии наук за 2005 год. Составители: Г.М.Михайлов, В.В.Дородницына, А.А.Муромский, Е.В.Панфилова, Н.П.Петрова, Н.П.Тучкова М.:ВЦ РАН
5. Моисеев Е.И., Муромский А.А., Тучкова Н.П. Проблема представления трудов научного коллектива в интернет с использование минимальных средств электронной среды //Труды Всероссийской научной конференции (22-27 сентября 2008 г., г. Новороссийск).- М.: Изд-во МГУ, 2008. - 468 с. (С. 370-374)
6. Большакова Е. И. О принципах построения компьютерного словаря общенаучной лексики // МГУ им. М.В. Ломоносова, факультет ВмиК.
7. Моисеев Е.И., Муромский А.А., Тучкова Н.П. Тезаурус информационно-поисковый по предметной области: обыкновенные дифференциальные уравнения. М.:МАКС Пресс, 2005, 116 с. ISBN 5-317-01489-1