О ТИПОВОЙ ЛИНГВИСТИЧЕСКОЙ МОДЕЛИ ИНТЕГРИРОВАННОЙ ОТРАСЛЕВОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ.

Маркарова Т.С.
(Научная педагогическая библиотека им. К.Д. Ушинского)

С учетом общемировых тенденций в области распределенного сетевого доступа к научной информации на базе сети Интернет, интеграционных процессов библиотек и архивов становится актуальным вопрос о разработке универсального лингвистического аппарата, кумулирующего в себе средства обработки разнородной информации. Однако до сих пор не только не найдено удовлетворительных решений этой проблемы, но даже отсутствует решение локальных задач, таких как корректная лингвистическая модель автоматизированной обработки архивных документов, коррелирующая с лингвистическим аппаратом библиотечного электронного каталога. Это связано как со спецификой контента различных систем и каталогов, так и с отсутствием универсального формата представления данных для самих электронных каталогов. Необходимо разработать схему такой типовой модели, которая позволила бы на своей основе создать распределенную лингвистическую базу, реализующую возможности поисковой навигации в базах данных библиотек и архивов. Это позволит перейти в будущем не только к сквозному поиску в электронном каталоге библиотеки и справочном аппарате архива, но и к практической разработке полнофункциональных, многопользовательских распределенных систем актуальной научной информации в её синхроническом и диахроническом срезе.

Качество функционирования любой информационной системы во многом зависит от уровня адекватного описания соответствующего фрагмента предметной области, достигаемого посредством соответствия информационной модели его (фрагмента) реальному состоянию. Для описания предметной области обычно используют комплекс

таких лингвистических средств, как естественные языки и искусственные формализованные языковые средства. Как правило, описание предметной области выполняется с помощью специальных языковых средств, не зависящих от используемых в дальнейшем программных средств. Основной задачей является получение формального (не зависимого от СУБД) описания предметной области, которая должна моделироваться. Методологическое значение моделирования для лингвистики определяется тем, что именно на принципах моделирования базируется практическая реализация методологических подходов к анализу основного объекта филологических наук — текста — и всех методов его исследования, основанных на использовании математического аппарата и средств вычислительной техники.

В результате моделирования создается промежуточный объект познания — модель, которая в процессе познания действительности выполняет ряд функций: замещения моделируемой системы (предметной области); информационную, гносеологическую; формализационно-алгоритмическую; доказательственно-иллюстративную и познавательную (модель в данном случае рассматривается как общенаучная категория, являющаяся системообразующей в процессе познания действительности). Перечисленные функции модели вытекают из следующих основных свойств:

- Абстрактность
- Наглядность
- Аналогичность
- Гипотетичность

Т.е., модель понимается как совокупность параметров, управляющих созданием, распространением, обработкой и использованием научной информации.

Моделирование выполняет функцию связующего звена между теорией и практикой. С одной стороны модель выступает в качестве вторичного объекта исследования, с другой, — как средство его фиксации. Моделирование представляет форму познавательной деятельности, базирующуюся на творческой отражательной способности субъекта, поэтому не может быть сведено к зеркальному

копированию изучаемого объекта или явления. Система целевых установок моделирования информации предопределяется онтологическими и гносеологическими признаками модели и предполагает:

- построение интерпретации информации с учетом референта (предмет реального мира) и корефферентов (средства обозначения и выражения);
- осмысление наглядно-оценочной характеристики модели;
- установление связей между моделью и экстралингвистической ситуацией;
- построение схематического образа модели как абстрактного представления связей структурно-семантических компонентов информации

При информационном моделировании приходится иметь дело с некоторыми сущностями, отсутствующими в тексте (форма представления информации) в явном формализованном виде, но несущими собственное реальное значение. В данном случае основной задачей становится выделение смысла информации в тексте, т.е. восстановление отдельных объектов и их взаимосвязей, которые либо описаны, либо упомянуты, либо подразумеваются неявно. Становится очевидным составление базы данных, в которой будет храниться информация о некоторых объектах, процессах и явлениях, описанных в текстах, и это записи должны сопровождаться информацией о свойствах, качествах и взаимосвязях описанных в текстах объектах. При этом один и тот же объект может описываться с использованием различных слов (терминов), либо даже не описываться, а упоминаться косвенно. Кроме того, в различных текстах (и даже нередко — в одном) одинаковыми словами могут описываться различные объекты (различные экземпляры, подклассы и т.п.). Важность решения данной проблемы в информационно-поисковой системе продиктована необходимостью, с одной стороны, сузить поиск, исключив из него документы, упоминающие ненужные пользователю объекты/события, с другой стороны — застраховаться от излишнего сужения, традиционно возникающего за счет того, что пользователь может спрашивать об объекте (персоне, собы-

тии) совсем не так, как это вербализовано в тексте. Функция восстановления объектов (событий) является самой простой для полноценной информационно-лингвистической поисковой системы. Вслед за проблемой восстановления объектов (событий) возникают проблемы восстановления взаимосвязей, отношений, характеристик и т.д. Но, если в первом случае задача решается известными средствами, такими как составление словарей объектов (событий), словарей-синонимов и т.д., то задача восстановления связей, отношений, характеристик, особенно описываемых неявно и/или разрозненно, усложнена тем, что не решается без применения интеллектуальных технологий, базирующихся на проработанных моделях естественного языка, моделях построения текстов, моделях мышления, математических моделях.

При использовании математического аппарата и средств вычислительной техники в создании информационно-лингвистических моделей система и методология проектирования должны поддерживать как знания о свойствах предметной области, так и отображение этих упорядоченных и организованных знаний в набор предварительных описаний, составляющих собственно информационную модель предметной области, выраженную в текстовом формате. Это предполагает решение ряда принципиальных задач, таких как:

- выстраивание за линейным текстом структуры контекста;
- создание базы данных целостного образа смысла текста;
- различение жанров текста;
- выделение из текста метатекстовых, коммуникативных и собственно содержательных составляющих.

Основными этапами разработки интегрированных информационно-отраслевых баз данных (БД) являются:

- формирование информационных массивов специализированных (отраслевых) БД из разнородных источников;
- организация информационных массивов специализированных БД для навигации с учетом заранее

установленных семантических (тезаурусных или онтологических) связей лексических единиц;
- совместимость специализированных БД с комплексом банков данных единой тематики;
- единое методическое обеспечение автоматизированных систем специализированных БД;
- единое лингвистическое обеспечение (использование комплекса лингвистического обеспечения интегрированной БД, классификаторов, словарей, справочников, тезауруса);
- прагматический аспект тематических БД, т.е. максимальное обеспечение практической ценности представляемой информации за счет представления не только отраслевой информации, но и полезной информации (справочной, энциклопедической, нормативной и т.д.);
- осуществление стратегии аналитического поиска (смысловой навигации в информационных пространствах специализированных БД) за счет установления тезаурусных, онтологических связей лексических единиц отраслевых понятий, используемых в аналогичных БД;
- отражение системности отраслевых понятий для создания семантической сети (семантически связанных понятий, отражаемых в тематических источниках разных областей отрасли). Для создания семантических сетей самым оптимальным является онтологический принцип обработки научной информации, т.е. структурная спецификация предметной области, ее формализованное представление, которое включает словарь (имена) указателей на термины предметной области и логические выражения, которые описывают, как они соотносятся друг с другом. Онтологии обеспечивают словарь для представления и обмена знаниями о некоторой предметной области и множество связей, установленных между терминами в этом словаре.

Стратегия аналитического поиска осуществляется по заранее определенным и установленным семантическим связям понятий. Поэтому распространение получили в

качестве готового продукта тематические БД с заранее установленными разработчиками (экспертами) маршрутами навигации, узлами-связями понятий. При этом предусматривается не только маршрутизация и навигация по смысловым связям понятий в застывшем виде, но и поддержание таких БД в динамичном, развивающимся состоянии с учетом пополнения узлов связи и возможностью быстрой модификации информационно-отраслевого продукта, поддержания его в актуальном состоянии.

Следует отметить, что на уровне библиографической базы данных, а именно, электронного каталога Учреждения Российской академии образования «Научной педагогической библиотеки им. К.Д. Ушинского» подобная модель уже функционирует. В 1998 году Лингвистической группой НПБ им. К.Д. Ушинского РАО была разработана модель связанного индексирования. В дальнейшем эта модель легла в основу процесса систематизации, который, в свою очередь, получил название «комплексное индексирование».

Суть данной модели заключается в том, что такие информационно-поисковые языки (ИПЯ) как ББК, УДК, ГрНТИ, DDC можно совместить на крупных иерархических уровнях. В качестве базового информационно-поискового языка может быть выбран один из них. В нашем случае это — ББК, т.к. по ней с 1987 года производится систематизация, строится систематический каталог библиотеки, и эта же система лежит в основе расстановки фондов.

Структура модели имеет следующий образ:

РУБРИКИ ПО ТЕМАМ:

Педагогика
Психология
Отдельные отрасли знания

РУБРИКИ ПО ВИДАМ:

Программы
Учебники
Конференции
Авторефераты диссертаций

Последняя рубрика вида повторяет перечень рубрик по темам (Педагогика, Психология, Отдельные отрасли знания).

Приоритетность тематико-видового выбора рубрик определена профильностью и спецификой НПБ им. К.Д. Ушинского РАО.

Рабочий лист каждой рубрики имеет следующие составляющие: раздел (или группа рубрик) и его план выражения через основные ИПЯ, рубрика, подрубрики, вышестоящая рубрика (для рубрик по видам), страна и соответствующие им ББК, УДК, ГРНТИ, DDC, ИИ (издательский индекс), ссылки «смотри» и «смотри также», фрагмент ББК для рубрик вида «Учебники», аннотация. Система предназначена для одновременного введения классификационных индексов вышеперечисленных ИПЯ в поисковый образ документа. Дробность индексов ББК соответствует делениям основных таблиц (ОТ) и плана расположения (ПР) — разделы «74 Педагогика» «88 Психология, Учебно-методические издания». Глубина индексов других отраслей знания ограничена, как правило, основными делениями (ОД). Однако, не исключается возможность продолжения индексов (ББК) до любого уровня, включая детализацию ПР, специальных типовых делений (СТД) и территориальных типовых делений (ТТД) в зависимости от специфики библиотек. Построение соответствия на уровне классификационных индексов идет, как уже говорилось, от ББК, индексы УДК, ГРНТИ, DDC выбраны по соответствию, если же соответствие установить невозможно, то совмещение устанавливается на уровне обозначения рубрики в целом.

Фрагмент модели текстового файла рубрики «Отдельные педагогические системы, школы, направления» в разделе «74 Педагогика»:

РАЗДЕЛ: Педагогика

ББК
74
УДК
37
ГрНТИ
14

DDC
370
ИИ
74

РУБРИКИ: Отдельные педагогические системы, школы, направления (наименование рубрики)

Россия (территориальный признак)

ПОДРУБРИКИ:

Отдельные педагогические системы, школы, направления в общей педагогике
Отдельные педагогические системы, школы, направления в дошкольной педагогике
Отдельные педагогические системы, школы направления в педагогике общеобразовательной школ
Отдельные педагогические системы, школы, направления в педагогике профессионального образования

Соответствующие планы выражения:

ББК:
74.003(2Рос)
74.103(2Рос)
74.203(2Рос)
74.500.3(2Рос)
УДК:
371.4(47)
ГРНТИ:
14.07.01
14.23.01
14.25.01
14.31.01&14.33.01&14.35.01
DDC(Дьюи):
370.04
ИИ(Издательский индекс):
74.003(2)

Каждая тематическая рубрика предваряется наименованием раздела (отрасли знания, например: «Педагогика»), которому принадлежит данная рубрика. Разделу

присваиваются соответствующие цифровые символы ББК, УДК, ГРНТИ, DDC. Сама рубрика представлена индексами ББК, ограниченными крупными делениями ПР. С каждым из индексов ББК связан конкретный индекс другой классификационной схемы: УДК, ГРНТИ, DDC. Индексы УДК и других схем выбраны по соответствию. Территориальные деления, которыми завершается тот или иной индекс выбраны следующие: по ББК для России — (2Рос), для зарубежных стран (3) без учета континентов и стран, эту функцию выполняет ИПЯ лексического типа (тезаурус), для УДК выбраны территориальные деления (47) и (100) соответственно. Будучи важным, но вторичным признаком, территориальный определитель в индексах DDC пока отсутствует, кроме тех случаев, которые отражают историю вопроса. Издательские индексы (ИИ) также входят в структуру модели и предназначены для создания указателя текущих поступлений в автоматизированном режиме. Причем, формирование БД указателя происходит одновременно с процессом систематизации. ИИ соответствуют первому индексу ББК (самый высокий уровень). ИИ действуют в пределах конкретной отрасли знания и идут в порядке возрастания цифровых обозначений. Это позволяет выстроить рубрики в заданной последовательности. Кроме того ИИ, являясь объединяющим значением для группы рубрик, позволяет разместить рядом близкие по значению, но находящиеся в разных отраслях знания понятия, как то: «социальные вопросы педагогики» и «социология образования», «организации и движения учащейся молодежи» и «молодежные организации и движения России или зарубежных стран» и т.п.

На базе этой модели было возможно и ведение карточного каталога при отсутствии глубокого индексирования по символическим ИПЯ, но при наличии дескрипторов тезауруса.

Механизм процесса индексирования по ББК сводится к правильному выбору индекса (одного или нескольких) в соответствующем разделе, рубрике. Каждый подлежащий смысловой обработке документ (библиографическая запись) подробно описывается дескрипторами Тезауруса согласно правилам индексирования, затем документу

присваивается нужный индекс ББК, классификационные индексы других схем присоединяются автоматически. При этом минимизируются ошибки при идентификации информации.

Процесс автоматизированного поиска информации в электронной БД (ЭБД) по различным ИПЯ происходит автономно, и каждый из языков несет свою функцию: тезаурус обеспечивает предметный вход в ЭБД; ББК служит, во-первых, инструментом организации знания (информации), во-вторых, предусматривает поиск по конкретному индексу, который, в свою очередь, обеспечивает системный подход в поиске данных; УДК, DDC, ГРНТИ несут ту же функцию, повышая при этом репрезентативность ЭБД Библиотеки — возможность входа по любому ИПЯ.

Однако, попытки адаптировать, приспособить эту модель к полнотекстовым базам данных даже идентичной предметной области пока не увенчались успехом, что, на наш взгляд, и доказывает необходимость разработки типовой лингвистической модели для полнотекстовых интегрированных баз данных, основанной на онтологических принципах построения семантической карты представления информации.

Литература:

1. *Барт Р. Избранные работы. Семиотика. Поэтика. — М.: Прогресс; Универс, 1994.*
2. *Винер Н. Информация, язык и общество // Кибернетика. — М.: Наука, 1983. С.236-248.*
3. *Волошинов В.Н. (М.М. Бахтин). Марксизм и философия языка: Основные проблемы социологического метода в науке о языке. — М.: Лабиринт, 1993.*
4. *Котов А.А. Семантические смещения в текстах СМИ и новые требования к когнитивным моделям//Вторая международная конференция по когнитивной науке: Тезисы докладов: В 2 т. — СПб: СПбГУ, 2006. — Т. 1. — С.322-323.*
5. *Матурана У. Биология познания // Язык и интеллект. — М.:Прогресс, 1995. С.95-142.*

6. *Розеншток-Хюсси О. Речь и действительность. — М.: Лабиринт, 1994.*
7. *Якобсон Р.О. Речевая коммуникация; Язык в отношении к другим системам коммуникации // Избранные работы. — М.: Прогресс, 1985. С.306-330.*
8. *McLuhan M. Essential McLuhan. N.Y.: Basic Books, 1995.*
9. *Shannon C. The Mathematical Theory of Communication // The Bell System Technical Journal. 1948. Vol.XXVII. # 3.*