УДК 81’27

Белоусов К.И., Зелянская Н.Л., Баранов Д.А.
Оренбургский государственный университет
E-mail: belousovki@gmail.com

КОНЦЕПТУАЛЬНО-ГИПЕРТЕКСТОВАЯ МОДЕЛЬ УПРАВЛЕНИЯ КОНТЕНТОМ В ИС «СЕМОГРАФ»

Статья посвящена описанию концептуально-гипертекстовой программы моделирования предметной области на основе информации о ней, репрезентированной в корпусе текстов. Построение концептуально-гипертекстовой модели, реализующей функцию управления контентом (текстами корпуса, ключевыми слова и их наборами, метаданными, понятийными полями и др.), осуществляется в информационной системе графосемантического моделирования «Семограф».

Ключевые слова: предметная область, онтология, ключевые слова, графосемантическое моделирование, ИС «Семограф».

Текстовая проекционность/воплощенность создания (и сам феномен его текстуальности), развитие информационных технологий, методов Datamining и Textmining, применяемых к анализу огромных текстовых массивов (например, для мониторинга социальных сетей), актуализируют проблему поиска, отбора, совершенствования методов релевантного предметным областям извлечения знаний из текстового материала. В лингвистике и других гуманитарных науках наблюдается большое оживление вокруг проблематики «онтостроительства» (моделирования предметных областей). Онтологическое моделирование предметных областей связано с построением тезауруса предметной области, охватывающего класс понятий и систему отношений между ними.

Под *предметной областью* мы понимаем аспект некоторой сферы (фрагмента) действительности, который выделяется, структурируется и интерпретируется в соответствии с целями, методами, инструментарием деятельности, осуществляемой над некоторым классом объектов очерченной сферы (фрагмента) действительности. Например, на классе объектов «тексты» мы можем выделить предметную область «психолингвистика текста», которая с помощью экспериментальных методов изучает механизмы порождения/восприятия/понимания текста (и текстов). Однако на том же классе объектов можно выделить и другие предметные области как внутри лингвистики и в смежных гуманитарных науках, так и в технических науках.

Онтологический анализ/моделирование предметных областей связан с построением тезауруса предметной области, охватывающего класс понятий и системы отношений между ними. При этом понятия, составляющие тезаурус предметной области, могут быть представлены в виде иерархической системы, каждый уровень которой описывает процесс и результат осуществления предметной деятельности над классом объектов действительности в соответствии с уровнями, выделяемыми в самой деятельности. Так, например, в процессе осуществления лингвистического анализа текста (ЛАТ) выделяются наиболее общие деятельностные (мыследеятельностные) операции анализа, синтеза, сравнения/сопоставления, обобщения, детализации и др.; деятельностные схемы, репрезентирующие общие подходы к научному анализу (например, диалектической логики, конкретно-исторического подхода, принципа взаимообусловленности формы и содержания; принципа уровневого анализа и др.); собственно филологические и/или лингвистические методы и приемы (метод компонентного анализа, семантико-стилистический метод, метод интертекстуального анализа, мотивный анализ и мн. др.) (подробнее см. [1; 4]). Данные уровни лингвистического анализа используют понятия, соответствующие каждому из уровней (например, понятия текст, автор, персонаж, мотив, лексико-семантическое поле относятся к разным уровням ЛАТ, причем некоторые могут присутствовать на нескольких деятельностных уровнях одновременно). Именно поэтому членение онтологии на онтологию верхнего уровня (базовую, общую), онтологию собственно предметной области, онтологию конкретной задачи (см., например, [12: 76]); или в терминах бизнес-проектирования: онтологии верхнего уровня, определяющего основные понятия организаций в общем, онтологии среднего уровня, определяющего по-

нятия специфичные для конкретной организации, и набора онтологий предметных областей знаний, в которых выполняется работа конкретной организации (см., например, [11: 138]), или онтологию базовую (тематическое ядро) [5: 97] и периферию; и мн. др. – представляют собой частные реализации множества возможных классификаций, осуществляемые в соответствии с целями моделирования, его достаточностью и полнотой для решения конкретных задач в сфере управления и принятия решений в данной предметной области.

Современный этап предметного моделирования немыслим без использования информационно-компьютерных технологий, функции которых зачастую состоят в генерации самих концепций моделирования предметных областей, в определении границ применимости концепций и во внедрении концепций в сферу практической деятельности. Сказанное относится к проблеме создания формализованных средств описания семантики данных в SemanticWeb (например, система формальных языков RFD/OWL, язык CycL для онтологической базы знаний или Standard Upperlevel Ontology (SUO), формализованных средств описания семантики предметных областей (например, языка Unified Modeling Language (UML), OilEd, основанный на языке OWL, а также простых в использовании программных средств, создаваемых в рамках концепции MindMap (MindMapper, Mindjet MindManage, FreeMind, Mind42.com и др.), использованию программных средств, созданных для моделирования предметных областей как на основе формализованных языков (например, IBM RationalRose и под. аналоги, используемые в основном для проектирования баз данных, а также в деятельности, связанной с анализом бизнес-процессов, систем управления и под.), так и с опорой на семантику естественного языка (например, InTez – онторедактор, основанный на универсальной словарной лексически интерпретированной онтологии; ResearchCyc – объемная онтологическая база знаний) и др. Одним из программных Web-приложений, созданным на основе современных технологий, которое можно использовать для моделирования предметных областей науки и практической деятельности, является информационная система графосемантического моделирования «Семограф» [1]. Информационная система (далее – ИС) графосемантического моделирования «Семограф» предназначена для извлечения знаний о предметных областях из информационных массивов, включающих текстовые выборки, метаданные, семантические компоненты и семантические поля, частотные, языковые и тезаурусные словари.

Благодаря реализации в ИС «Семограф» принципа универсальности – применимости ИС для решения широкого спектра задач, связанных с экспертным анализом текстовой информации, моделирование предметных областей может осуществляться разными способами. В то же время можно предложить наиболее полное описание предметной области, осуществляемое в рамках разрабатываемой нами исследовательской программы, получившей название концептуально-гипертекстового моделирования КГМ(у), описание которой приводится ниже.

Обобщенное описание этапов программы моделирования предметной области

Предлагаемая исследовательская программа изучения частнонаучных (узкопредметных) областей реализуется в рамках антропоцентрического подхода к моделированию онтологии предметной области, т. е. осуществляется на материале корпуса научных текстов (не словарей!), относящихся к данной предметной области. Сама исследовательская программа состоит из нескольких этапов.

На первом этапе создается тезаурус предметной области с учетом частотности используемых понятий. Частотность понятийного аппарата позволит определить ядро онтологии предметной области, ее периферию, единичные использования понятий. Также появляется возможность сопоставления онтологий разных предметных областей, что позволит определить единицы, которые входят в

[1] Информационная система графосемантического моделирования «Семограф» (http://semograf.com) создавалась как пилотный проект и начала функционировать в 2010 году [9]. Несмотря на отсутствие какой бы то ни было рекламы, в ИС (информационной системе) на сегодняшний момент зарегистрировано 176 пользователей, создано 243 проекта, внесено в базу данных 42984 контекста, выделено 1262 поля и 17858 компонентов. Эти данные являются свидетельством востребованности системы, того инструментария, который предлагается исследователю. В настоящее время готовится промышленный запуск новой версии ИС «Семограф», которая полностью заменит старую пилотную версию.

ядро одних предметных областей и в периферийные области в других онтологиях, т. е. установить влияние одних онтологий на другие. Кроме того, на данном этапе представляется возможным выявить совместную встречаемость двух понятий предметной онтологии в научных текстах, т. е. определить концептуальную близость двух понятий.

На втором этапе на материале всего тезауруса предметной области строится ее обобщенная информационная модель. Также выявляются зависимости между временными (дата публикации), топологическими (географическими) и социальными факторами и моделируемым информационным пространством предметной области. Созданные модели научных онтологий позволят оценить предметную область на шкалах междисциплинарности/монодисциплинарности, лабильности/ригидности, «обновляемости» проблемного поля наук, что одновременно будет являться и характеристикой научного мышления, сложившегося в данных предметных областях.

На третьем этапе моделируется понятийный потенциал ядра онтологии предметной области. Выявленные на первом этапе наиболее значимые (значимость определяется частотой употребления) понятия сами становятся предметом исследования. Так как данные понятия и соответствующие им термины часто используются в исследуемой предметной области, возникают уже описанные выше явления концептуализации термина, приращения новых смыслов, осмысления его в новых контекстах, интерпретация посредством новых концептуальных структур. И поэтому на третьем этапе производится моделирование понятийного потенциала ядра онтологии предметной области с опорой как на дефиниции терминов, так и на более широкие контексты их употребления.

На четвертом этапе моделирования научной картины мира (НКМ), осуществляемом в ИС «Семограф», создается концептуально-гипертекстовая модель управления корпусом текстов, используемых в исследовании.

Последовательность этапов реализации программы моделирования НКМ представлена на рисунке 1.

На рисунке 1 показано, что после экспликации ядра онтологии предметной области можно переходить к созданию гипертекстовой модели управления корпусом текстов, используемых в исследовании. Недостатком такого подхода является то, что создаваемая модель управления будет работать не на всем корпусе текстов, а только на той ее части, которая маркируется ядром онтологии предметной области. Тексты же, маркируемые нечастотными понятиями предметной области не смогут быть проиндексированы. В то же время поэтапная реализация программы (1 этап – 2 этап – 4 этап) позволит работать со всей предметной областью и, соответственно, всем корпусом текстов. Третий этап реализации программы служит целям, не связанным непосредственно с созданием концептуально-гипертекстовой модели управления контентом.

На четвертом этапе моделирования предметной области, осуществляемом в ИС «Семограф», создается концептуально-гипертекстовая модель управления (КГМ(у)) корпусом текстов.

Понятие гипертекста разрабатывается в современных гуманитарных науках по нескольким направлениям:

1) гипертекст как «набор» текстов, между которыми существуют материально выраженные правила перехода в форме ссылок/гиперссылок (электронный гипертекст, научная энциклопедия, варианты художественного гипертекста и т. п.);

2) гипертекст, в котором гипертекстуальность следует понимать как способ управления чтением (смыслообразованием, построением «читательской проекции»), основанный на знании широкого культурного контекста – здесь отсылки к другим текстам не выражены в явной форме, они аллюзивны;

3) гипертекст как выборка текстов (в семиотическом понимании термина текст) в том случае, если смоделирована такая пред-

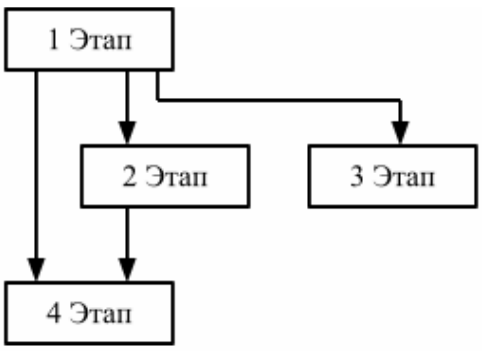
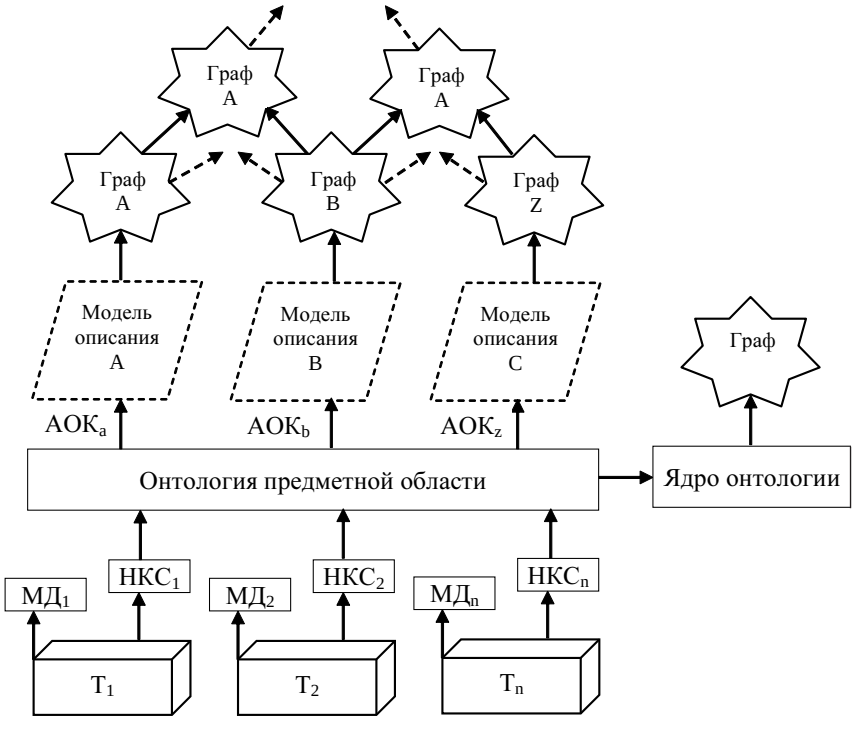


Рисунок 1. Этапы реализации программы моделирования предметной области

метная область, в которой установлены правила связности между текстами этой совокупности, а онтология предметной области представлена в виде концепта/концептосферы со связной системой смысловых/понятийных полей, в свою очередь, репрезентированных в текстах совокупности. *В качестве гиперссылочного инструментария выступает пользовательский интерфейс, созданный под задачи управления фильтрацией, поиском по семантическому графу онтологии предметной области и выводом статистических результатов с их визуальным представлением.*

В качестве совокупности текстов могут выступать самые разнообразные текстовые массивы, в частности, газетный текст в совокупности всех своих статей [8], совокупность текстов-граффити на парте [10], выборка определений термина [2], совокупность коммерческих номинаций для товаров и услуг [3]; совокупность заглавий прозаических произведений эпохи [6]; выборка медиатекстов, формирующих медиаобраз [7] и мн. др.

Новым в понимании природы гипертекста является третий из описанных типов. Именно для данного типа гипертекста важнейшим компонентом создаваемой теоретической схемы становится связь между гипертекстом и концептом, представленным в виде связной семантической модели: концепт (концептосфера) является средством обеспечения связности текстов, их структурации, иерархизации, т. е. средством создания самого гипертекста. Если проводить аналогию отношений между гипертекстом и концептом, то это будет система отношений между телом текста и его проекцией. В нашем случае концепт – это структурированная семантическая проекция гипертекста, а сам гипертекст – материальная основа существования данной проекции. Данный методологический постулат позволяет продемонстрировать потенциал самоорганизации текстов, сосуществующих в качестве самостоятельных, но тесно связанных семиотических фактов, обнаружить логику структурирования объединяющего их смыслового пространства (логику построения гипертекста).



Т1, Т2,..., Тn – пронумерованные тексты в корпусе;
НКС1, НКС2,..., НКСn – наборы ключевых слов для каждого текста;
МД1, МД2,..., МДn – метаданные, характеризующие каждый текст;
АОКa, АОКb,..., АОКz – альтернативные описания контекстов;
НАКi,..., НАКs – наборы активных контекстов.

Рисунок 2. Концептуально-гипертекстовая модель управления контентом

На основе такого понимания соотношения гипертекста и репрезентирующей его онтологической семантики строится концептуально-гипертекстовая модель управления контентом (см. рисунок 2). На рисунке 2 видно, что модель строится на корпусе текстов Т1, Т2, Т3,..., Тn. Корпус текстов может постоянно пополняться с помощью автоматизированного мониторинга предметной области, осуществляемого отдельным модулем ИС. После такого рода изменений автоматически производится пересчет частотности понятий, составляющих онтологию предметной области, определяется ядро онтологии, автоматически вычисляется семантическая карта, отражающая совместную встречаемость понятий в корпусе текстов, и строится се-

мантический граф, визуально репрезентирующий структуру отношений, определенных в семантической карте. Временные изменения в структуре ядра онтологии позволяют вскрыть логику развития предметной области.

Каждый текст (в терминах ИС – контекст), как уже отмечалось выше, описывается системой метаданных (МДr) и компонентов, в качестве которых в нашем исследовании выступают ключевые слова к статьям (НКСr). Совокупность НКС всего корпуса, рассматриваемых в качестве понятийно-категориального аппарата предметной области, составляют ее онтологию. На основе частотного критерия можно выделить ядро онтологии предметной области (и другие ее зоны) и представить в виде графа структуру отношений между базовыми понятиями предметной области.

Но, как уже отмечалось, граф, характеризующий отношения между базовыми понятиями предметной области, способен индексировать небольшую часть текстов корпуса. Поэтому возникает необходимость в осуществлении второго этапа программы моделировании предметной области – классификации и группировки понятий. В ИС «Семограф» такие группы носят наименование полей.

Однако можно предположить, что результаты полевого анализа – набор полей – могут отличаться в зависимости от инструментов классификации. На рисунке 2 этому аспекту работы с предметной областью соответствуют «Модель описания А», «Модель описания В» и «Модель описания С». Возможность осуществления нескольких вариантов/моделей описания одного и того же материала реализуется в ИС «Семограф» посредством концепции альтернативных описаний контекстов (АОК).

Совместное использование инструментов НАК (набора активных контекстов) и АОК [2] технологически осуществляется следующим образом. В модели описания АОКs предметной области осуществляется фильтрация корпуса текстов по значениям полей (не по метаданным!). Например, отбираются все контексты, которые характеризуются полями КОНЦЕПТ И КОНЦЕПТОСФЕРА и МОДЕЛЬ И МОДЕЛИРОВАНИЕ. При этом каждое из полей состоит из множества понятий. В результате из корпуса текстов сформируется выборка Y с набором активных контекстов НАКi. С данной выборкой Y можно в пределах этой же модели описания получить много самых разнообразных моделей, связывающих семантику полей со значениями метаданных, например, установить, с какими вузами и/или в какое время и/или с каким научным статусом соотносятся исследования в данной области (моделирование концептов/концептосфер) и т. п. Полученные результаты автоматически представляются в виде таблиц и графиков и могут использоваться для «умного» поиска контента.

Однако сформированную выборку Y с набором активных контекстов НАКi можно использовать для работы с альтернативной моделью описания АОКq. Для этого набор номеров НАКi «передается» в модель описания АОКq, в которой из всего корпуса активируются только те контексты, которые были активированы в АОКs, т. е. НАКi. После формирования выборки Y с набором активных контекстов НАКi в модели описания АОКq вычисляется семантическая карта, граф, таблицы частотности, осуществляется анализ метаданных, соотносящий поля АОКq со значениями метаданных этой выборки МДi. Таким образом, данная модель реализует инструментарий для соотнесения как экспертных моделей описания предметных областей, а также и вообще любых моделей описания одного и того же корпуса текстов (например, моделей, созданных в результате тематического и стилистического анализа текстов корпуса).

Кроме режима «ручного» (выборочного, мотивированного экспертным суждением) соотнесения двух моделей описания корпуса текстов возможно автоматизированное построение в любой модели описания АОК = {АОК1, АОК2,..., АОКm} графов, характеризующих наиболее вероятные связи структуры полей в какой-то конкретной модели из той же системы моделей АОК. То есть вероятностное состояние системы одного описа-

[2] Напомним, что инструментарий НАК на основе значений метаданных (и полей) позволяет осуществлять фильтрацию всего корпуса текстов и составлять из него выборки. Однако инструментарий НАК функционирует в пределах одной модели описания контекстов. Поэтому использование АОК значительно расширяет возможности функционала Семографа при моделировании предметной области.

ния может быть оценено в другой системе описания тоже с позиций вероятности его появления. Данные возможности концептуально-гипертекстовой модели реализуются посредством пользовательского интерфейса с функциями поиска, фильтрации, множественной и вложенной сортировки и генерации статистических таблиц, графиков и графов по результатам работы с корпусом в данном интерфейсе.

16.09.2012

Исследование выполнялось при финансовой поддержке Российского гуманитарного научного фонда (проект № 12-04-12034в)

Список литературы:

1. Белоусов, К.И. Филологические модели: теоретические и прикладные аспекты филологических исследований / К.И. Белоусов, Н.Л. Зелянская. – Saarbrucken: LAP Lambert Academic Publishing, 2011. – 224 с.
2. Белоусов, К.И. Моделирование понятийного потенциала термина заглавие / К.И. Белоусов, Н.Л. Зелянская // Известия высших учебных заведений. Поволжский регион. Гуманитарные науки. – Пенза, 2008. – С. 62–71.
3. Белоусов, К.И. Имя для пельменей (мониторинг ассоциативно-смысловых ожиданий потребителей) / К.И. Белоусов, Н.Л. Зелянская // Маркетинг в России и за рубежом. – 2007. – №1. – С. 11–19.
4. Головина, Е.В. Терминологический тезаурус филологического анализа текста как поле интерпретаций / Е.В. Головина // Речеведение: современное состояние и перспективы. – Пермь: Изд-во Перм. гос. ун-т, 2010. – С. 478–483.
5. Данилова, В.С. Современные проблемы дисциплинарных онтологий (физика, техника) / В.С. Данилова, П.П. Кожевников // Вестник Северо-Восточного федерального университета им. М.К. Аммосова. – 2007. – Т. 4, №1. – С. 97–105.
6. Зелянская, Н.Л. Культурно-семиотические тенденции русской прозы 1850-х годов / Н.Л. Зелянская // Сибирский филологический журнал. – 2008. – №4. – С. 40–51.
7. Зелянская Н.Л., Гавенко А.С., Белоусов К.И. Медиа-образ Иосифа Сталина как гипертекст // Вестник Оренбургского государственного университета. – 2009. – №11. – С. 73–79.
8. Зелянская, Н.Л. Языковая репрезентация политического компонента картины мира в заголовках СМИ / Н.Л. Зелянская, Ж.А. Никифорова // Вестник Оренбургского государственного университета. – 2011. – №11 (130). – С. 111–113.
9. Система графосемантического моделирования [электр. издание] / Д.А. Баранов, К.И. Белоусов, И.В. Влацкая, Н.Л. Зелянская. – М.: Свидетельство о государственной регистрации в Федеральной службе по интеллектуальной собственности, патентам и товарным знакам. Зарегистрировано в Реестре программ для ЭВМ № 20111617192 от 15.09.2011.
10. Стренева, Н.В. Композиционно-графический фрейм текста (на материале граффити): автореф. дис. канд. филол. наук / Н.В. Стренева. – Уфа: Изд-во БашГУ, 2009. – 21 с.
11. Тузовский, А.Ф. Метод объединения онтологий предметных областей знаний / А.Ф. Тузовский // Известия Томского политехнического университета. – 2006. – Т. 309, №7. – С. 138–141.
12. Ханова, А.А. Предметная онтология как способ формирования семантической модели знаний грузового порта // А.А. Ханова, И.О. Григорьева // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. – 2009. – №1. – С. 76–81.

Сведения об авторах:

Белоусов Константин Игоревич, профессор кафедры электронных СМИ Оренбургского государственного университета, доктор филологических наук
460018, г. Оренбург, пр-т Победы, 13, ауд. 1106б, e-mail: belousovki@gmail.com

Зелянская Наталья Львовна, ведущий научный сотрудник лаборатории филологического моделирования и проектирования Оренбургского государственного университета, кандидат филологических наук
460018, г. Оренбург, пр-т Победы, 13, ауд. 1106б, e-mail: zelyanskaya@gmail.com

Баранов Дмитрий Александрович, аспирант кафедры математического обеспечения информационных систем Оренбургского государственного университета
460018, г. Оренбург, пр-т Победы, 13, ауд. 2131, e-mail: demon.asgard@gmail.com

UDC 81'27

Belousov K.I., Zelyanskaya N.L., Baranov D.A.
Orenburg state university, e-mail: belousovki@gmail.com

CONCEPTUAL AND HYPERTEXT MODEL OF CONTENT MANAGEMENT IN THE SYSTEM «SEMOGRAF»

The article describes the conceptual hypertext software application of domain modeling based on the information about it represented in a corpus of texts. Building a conceptual hypertext model realizing the function of content management (text body, key words and their sets, metadata, conceptual fields, etc.), is carried out in an information system of graphosemantic modeling «Semograf».

Key words: subject area, ontology, key words, grafosemantic modeling, IS «Semograf».