УДК 004.82:[002:004]

И. М. Зацман

Вербально-образное представление знаний в электронных библиотеках*. Ч. I

Рассматривается концепция вербально-образного представления знаний и семиотической аппроксимации графических данных как теоретическая основа семантического моделирования и поиска вербальной и невербальной информации в политематических электронных библиотеках научных документов. Основное внимание уделяется принципам построения знаковых систем для организации семантического поиска. Для представления знаний и организации поиска в электронных библиотеках предлагается вербально-образный тезаурус, являющийся основой построения мультимодальных семиотических систем и включающий вербальные, вербально-образные и образные дескрипторы.

1. ВВЕДЕНИЕ

Актуальность проблемы семантического моделирования и поиска вербальной и невербальной информации значительно возросла с появлением предметно-ориентированных и политематических электронных библиотек, интегрирующих большие объемы научных информационных ресурсов в цифровой форме. Появились электронные коллекции и библиотеки научных документов, включающих, кроме вербальной информации, математические и структурные химические формулы, таблицы, карты, схемы, рисунки, диаграммы. Кроме того, в компьютерные документы авторы иногда включают анимационные, пространственно-временные и потоковые информационные объекты (аудио и видео).

Теоретические и прикладные исследования в информатике, направленные на решение фундаментальной проблемы семантического поиска, ведутся на основе двух концептуально противоположных направлений. Сторонники первого направления считают вербальную информацию документа первичной с точки зрения представления знаний, а невербальную — вторичной [1, 2]. Т. е. аксиоматически предполагается, что научные знания отражены в основном через вербальные коммуникативные компоненты научных документов. В рамках этого направления задачи семантического моделирования и поиска научной информации ставятся и решаются не в полном семантическом пространстве электронной библиотеки, а только в вербальной области этого пространства.

Сторонники второго направления считают невербальную научную информацию не менее ценной, чем вербальную [3, 4]. Более того, утверждается, что значительная часть научных знаний может быть выражена только в невербальной форме. Этому представлению научных знаний и их пониманию, в том числе в процессе обучения, в последние годы посвящен ряд работ по семиотике научной информации и проблемам образования [5–7]. Результаты этих работ показывают, что для решения проблемы семантического моделирования и поиска необходимо учитывать содержательные аспекты и вербальных, и невербальных компонентов научных документов [8], а также три основных способа передачи смысла в документах: презентационный, организационный и ориентационный [5].

Если сравнивать разные науки и предметные области знаний, то при обработке геодокументов, относящихся к наукам о Земле, доминирует второе направление. Для специалистов в этой области знаний очевидно, что содержательные аспекты топографических и тематических карт и схем, аэрофотоснимков, литологических и стратиграфических колонок, картоподобных диаграмм, палеотектонических схем и других геоизображений, являющихся составной частью геодокументов, невозможно адекватно передать в вербальной форме.

Семантический поиск научной информации в электронных библиотеках, содержащих геодокументы, должен учитывать содержательные аспекты и вербальных, и невербальных компонентов [9]. При этом содержательные аспекты невербальной информации не всегда могут быть переданы через вербальные метаописания. Последнее утверждение является частным случаем ключевого положения теории семиотики о трех основных сферах представления знаний: невербальные знания, которые не могут быть представлены в лингвистической форме, вербальные знания, которые не могут быть адекватно переведены в невербальную форму, и та часть знаний, которая может быть представлена и в вербальной, и в невербальной формах [10].

В настоящее время одним из способов представления в электронных библиотеках невербальной информации является кодирование ее содержательных аспектов с помощью метаданных в литерной форме на основе некоторого стандарта. Так представляются описания карт и геообъектов. Метаданные карт включают информацию о тематической направленности, содержании, объеме, точности и других характеристиках карт. Метаданные

* Работа выполнена при частичной поддержке РФФИ в рамках проекта № 01-06-80332.

геообъектов состоят из структурированной совокупности полей, содержащих сведения о пространственной организации и локализации геообъектов, их атрибуты, временные характеристики, которые могут включаться в запросы на поиск геоданных в электронных библиотеках [11, 12].

С одной стороны, благодаря этому можно расширить пространство семантического поиска за счет тех содержательных аспектов карт и геоизображений, которые могут быть достаточно адекватно переданы в литерной форме. С другой стороны, в силу отмеченного выше основного положения семиотики, при использовании метаописаний для семантического поиска остается закрытой сфера невербального представления знаний, для которой отсутствуют адекватные вербальные формы.

В рамках второго направления был предложен новый класс логико-семантических моделей научных документов, ориентированных на решение проблемы семантического поиска, а также перечень основных этапов логико-семантического моделирования корпуса полнотекстовых научных документов [13]. Закономерность перехода от логико-лингвистических моделей к логико-семантическим рассмотрена в работе [14].

В работе [13] предлагалось на первом этапе моделирования проводить декомпозицию документа на однородные и неоднородные коммуникативные компоненты, а на второй — получать знаковые формы представления компонентов в процессе их семантической разметки и кодирования. Основа реализации первого этапа — известные схемы декомпозиции (иерархическая, сетевая, реляционная, параметрическая, пространственно-временная) и их сочетания, а второго — системы знаков, которые можно было бы использовать в процессе семантического кодирования однородных и неоднородных коммуникативных компонентов.

Однако в настоящее время только для вербальных и отдельных видов структурных компонентов (например, структурные химические формулы) определены, исследованы и используются соответствующие системы знаков. Следовательно, сейчас только отдельные виды коммуникативных компонентов документов могут быть представлены в знаковой форме. Проблема знакового представления всех вербальных и невербальных компонентов научных документов в электронных библиотеках не решена.

Если отвлечься от электронных библиотек и обратиться к семиотике, то "общая семиотическая теории единого информационного мира еще даже не начала создаваться" [15]. Поэтому, в отсутствие общей теории, предлагается ограничиться рассмотрением семиотических характеристик коммуникативных компонентов документов электронных библиотек. И только в пределах библиотек, как относительно замкнутых образований, по сравнению с единым информационным миром, попытаться определить основные семиотические характеристики вербальных и невербальных компонентов, а также принципы и методы построения систем знаков для вербально-образного представления знаний в электронных библиотеках. При этом будем исходить из ориентации построения систем знаков на решение проблем поиска вербальной и невербальной информации во всем семантическом пространстве электронной библиотеки.

2. ОСНОВНЫЕ ТЕРМИНЫ И ПОНЯТИЯ

Прежде чем перейти к изложению основных положений предлагаемой концепции вербально-образного представления знаний, приведем определения терминов и понятий, используемых в статье.

Вербальные компоненты документов — монолинейные дискретные конкатенации литер, детерминированных по своей форме и очертаниям, к которым относятся естественно-языковые фрагменты названия, аннотации, разделов, глав и параграфов, подрисуночных подписей, а также текст на естественном языке на диаграммах, картах, схемах и графиках, в ячейках таблиц. Иногда в текстах на естественных языках встречаются фрагменты с полилинейной дискретной конкатенацией за счет использования литер "титло", "лига", "прогрессивная ассимиляция", "регрессивная ассимиляция", "гиперфонема", а также литер стрелок, указывающих на высокий/низкий уровни тона и движение тона с низкого уровня на высокий, и наоборот, которые располагаются над словами [16]. Такие фрагменты будем также относить к вербальным компонентам.

Структурные компоненты — полилинейные дискретные конкатенации литер, знаки с детерминированной формой, а также совокупности литер и знаков, которые связаны с помощью сетевых, иерархических, реляционных, дискретных параметрических схем и их сочетаний. К ним относятся математические формулы, структурные химические формулы и реакции, биоинформационные последовательности и таблицы.

Графические компоненты — континуальные или дискретно-континуальные сочетания знаков, формы которых могут быть одномерными и многомерными, статичными и динамичными, детерминированными, размытыми, случайными и неопределенными. К таким компонентам относятся графики, диаграммы, схемы, чертежи, карты, рисунки и фотографии, за исключением вложенных вербальных и структурных компонентов, а также динамические информационные объекты (анимационные, потоковые и пространственно-временные) в компьютерных документах.

Описание упомянутых в определениях схем организации документов и их компонентов дано в работах [9, 13].

Компоненты документов могут быть многократно вложены друг в друга. Например, таблица может иметь текстовые ячейки, ячейки со структурными химическими формулами или рисунками. Сами таблицы могут быть многократно вложены друг в друга. За счет многоуровневой вложенности и/или сочетания *однородных компонентов* (вербальные, структурные, графические) могут быть получены четыре вида *неоднородных компонентов* (вербально-структурные, вербально-графические, структурно-графические и вербально-структурно-графические).

Для обозначения всех видов однородных и неоднородных компонентов используется термин *коммуникативные компоненты документа*, а для обозначения структурных, графических и любых видов неоднородных компонентов — *невербальные компоненты*.

Традиционный документ определим как совокупность вербальных и/или невербальных компонентов с детерминированными по форме, статичными, одномерными или двумерными знаками. *Обобщенный документ* (generalized document) включает компоненты с размытыми, случайными или неопределенными по форме знаками, которые могут быть многомерными и динамичными [17].

Семантический поиск определяется как поиск по содержательным аспектам всех компонентов документов электронной библиотеки, включая три основных способа передачи смысла в документах: презентационный, организационный и ориентационный.

Случай, когда пользователю не известна компонентная форма представления в электронной библиотеке интересующих его сведений (вербальная, структурная, графическая, вербально-структурная, вербально-графическая, структурно-графическая или вербально-структурно-графическая), определим как *обобщенный семантический поиск*.

Если в научном документе авторы определяют значения знаков, которые используются только ими, то такие знаки будем называть *авторскими*. Знаки, форма и значение которых сохраняют общепринятое единство, устойчивое в течение достаточно длительного периода времени, будем называть *определенными*. Введем понятие *слабоопределенных знаков* для обозначения тех случаев, когда используемые знаки не являются общепринятыми, но при этом они используются достаточно большим числом авторов.

Когда в статье речь идет о научных документах, то имеются в виду и традиционные, изначально созданные на бумаге документы, и обобщенные документы, создаваемые изначально в цифровой компьютерной форме.

3. ЗНАКОВОЕ ПРЕДСТАВЛЕНИЕ КОМПОНЕНТОВ НАУЧНЫХ ДОКУМЕНТОВ

В работах [8, 9, 13] рассмотрена целевая ориентация логико-семантического моделирования корпуса научных документов на решение проблемы семантического поиска в электронных библиотеках научной информации, в том числе и в геобиблиотеках. Определен перечень основных схем структуризации документов и основные стадии построения логико-семантических моделей.

На **первой стадии** осуществляется декомпозиция документа на однородные и неоднородные компоненты на основе схем структуризации и с учетом возможной многоуровневой вложенности компонентов. В результате мы получаем логическую модель документа с указанием всех использованных схем структуризации, адресуемые в соответствие с этими схемами однородные и неоднородные компоненты документа, а также кодированные семантические связи между компонентами. В настоящее время разработка методов для реализации этой стадии моделирования является предметом целого ряда прикладных исследований и разработок (см. обзор в [18]).

Цель **второй стадии** — получение знаковых представлений всех компонентов документов в процессе их семантической разметки и кодирования с использованием систем вербальных и невербальных знаков. Эти системы знаков предлагается строить на основе введенной типологии вербальных и невербальных компонентов. Эта стадия моделирования состоит из двух этапов. Основным содержанием *первого этапа* является получение знаковых представлений компонентов, которые являются семиотической основой для вербально-образного представления знаний и организации семантического поиска сведений в электронных библиотеках.

Цель *второго этапа* — формирование семантического пространства электронной библиотеки с помощью тезауруса на основе корпуса документов, представленных в знаковой форме. При этом предполагается использовать знаковые представления компонентов, полученные на основе совокупности систем вербальных и невербальных знаков, которую будем называть *семиотической системой* электронной библиотеки. Пока не определены принципы ее построения и не построена сама семиотическая система электронной библиотеки невозможно получить и знаковые представления всех компонентов ее документов.

На сегодняшнем этапе развития электронных библиотек только вербальные и отдельные виды структурных компонентов могут быть представлены в знаковой форме. Используемые в настоящее время языки семантической разметки и кодирования полнотекстовых документов, как правило, на них и ориентированы. Проблема знакового представления всех вербальных и невербальных компонентов научных документов, включая широкий спектр графической информации в электронных библиотеках, не решена [19].

Для знакового представления графических компонентов требуется существенное развитие языков семантической разметки, в первую очередь, для различных видов континуальных и дискретно-континуальных изображений (карт, графиков, диаграмм, схем, чертежей, рисунков и фотографий), в которых частично или полностью отсутствует дискретная отделимость знаков, характерная для вербальных компонентов документов.

Рассмотрим проблемы знакового представления и их место в процессе логико-семантического моделирования на примерах неоднородных компонентов научных документов.

Первая стадия моделирования — декомпозиции документа — включает два этапа его структуризации. На первом этапе выбирается базовая схема описания структуры всего документа. Для коллекции документов с относительно регулярной структурой первый этап декомпозиции может выполняться один раз для всей коллекции или для всех документов одного типа. Такой случай построения единой базовой схемы для всех документов одного типа рассмотрен в работе [13] для коллекции полнотекстовых отчетов по инициативным проектам Российского фонда фундаментальных исследований. Однако компоненты, получаемые на первом этапе декомпозиции, могут иметь вложенные компоненты разных видов, т. е. являться неоднородными компонентами.

Например, график, полученный в результате первого этапа декомпозиции статьи [20] и изображенный на рис. 1, является неоднородным компонентом.

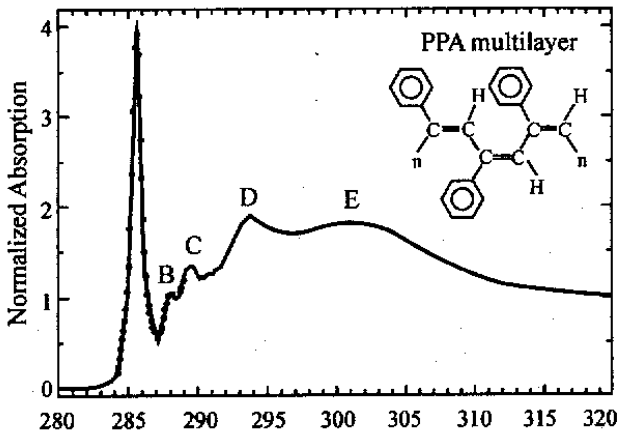

Рис. 1. Структурная химическая формула и вербальные компоненты графика

Этот график включает вербальные компоненты ("Normalized Absorption", "PPA multilayer"), пять символов ("A", "B", "C", "D" и "E", которые можно рассматривать как вырожденные вербальные компоненты из-за отсутствия конкатенации) и структурную химическую формулу. Однако в базовой структурной схеме документа эти вербальные компоненты и структурная химическая формула не выделяются в виде вложенных в график и отдельно адресуемых компонентов документа.

Цель второго этапа декомпозиции — структурно выделить адресуемые однородные компоненты, используя необходимое для решения задач семантического поиска количество уровней структуризации и сочетаний схем описания связей между компонентами. Исследование потенциальных возможностей семантического поиска в зависимости от детальности декомпозиции документов, в частности от числа уровней структуризации, представляет самостоятельную актуальную проблему, которая здесь не рассматривается.

В качестве иллюстрации второго этапа декомпозиции рассмотрим компоненты, приведенные на рис. 1. Химическая формула, изображенная в правом верхнем углу рисунка, может быть выделена на втором этапе как структурный компонент документа. В этом случае, сочетание графика, вербальных компонентов и структурной химической формулы идентифицируется как неоднородный компонент документа (вербально-структурно-графический), имеющий внутреннюю структуру и разнотипные адресуемые компоненты.

Вторым примером неоднородного компонента может служить изображение места впадения реки в озеро (рис. 2). Этот пример вербально-графического компонента из книги "Язык карты" [21, с. 116] включает три вербальных компонента ("пос. Верхний", "пос. Нижний" и "Озеро"), которые могут быть выделены, идентифицированы как отдельно адресуемые компоненты и использованы при вербальном поиске.

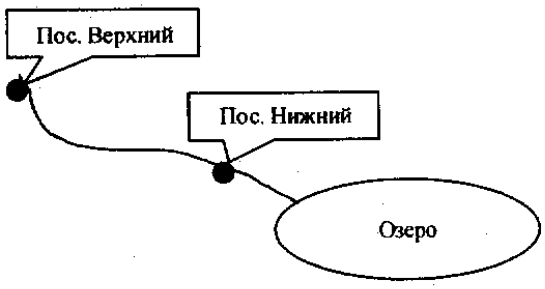

Рис. 2. Изображение течения реки

При выборе глубины декомпозиции подобных вербально-графических компонентов возможны следующие варианты декомпозиции: 1) с выделением подписи под рисунком, но без выделения вербальных компонентов и с адресацией единого вербально-графического компонента без подписи и 2) с выделением подписи, вербальных и однородного графического компонентов.

По содержательным аспектам эти варианты декомпозиции отличаются. Второй вариант декомпозиции дает графический компонент, который отображает более общий концепт, представленный только графической формой. По смыслу этот компонент охватывает два частных случая течения реки, а именно, "впадающей в" и "вытекающей из". Первый вариант отображает более частный концепт, соответствующий только случаю впадающей реки.

На второй стадии логико-семантического моделирования эти два варианта декомпозиции могут иногда давать один и тот же результат в семантическом пространстве электронной библиотеки, например, когда тезаурус электронной библиотеки отражает оба упомянутых концепта и содержит соответствующее родо-видовое отношение между ними.

На второй стадии моделирования рассматривается совокупность документов, предназначенных для создания новой электронной библиотеки или для пополнения уже существующей. Основная цель второй стадии — получить знаковое представление электронных форм всех вербальных и невербальных компонентов, а также сформировать семантическое пространство электронной библиотеки. Однако получение знаковых представлений графических и неоднородных компонентов документов является достаточно сложной задачей.

Для вербальных и отдельных классов структурных компонентов знаковые представления можно получить, используя слова и устойчивые словосочетания естественных языков, а также языки семантической разметки и кодирования структурных компонентов. Например, получив на стадии декомпозиции структурную химическую формулу, для ее знакового представления можно использовать язык семантической разметки структурной химической информации CML (Chemical Markup Language). Существуют также аналогичные языки для семантической разметки математических выражений и биоинформационных последовательностей [19].

Для подавляющего числа графических и неоднородных компонентов осуществить семантическую разметку и получить их знаковые представления достаточно трудно. Большинство рисунков, карт и других континуальных изображений не имеет однозначного, дискретного и детерминированного квантования на составляющие их знаки и сочетания знаков. Как и в случае вербальных компонентов, для графики существует проблема нормализации графических знаков. Но для графических компонентов проблему нормализации приходится рассматривать в сочетании с проблемой многозначности и возможной недетерминированности квантования графических компонентов на знаки и их сочетания.

Для вербальных компонентов, как правило, достаточно просто установить языковую принадлежность вербальных знаков. Для графических компонентов ситуация качественно иная. Как будет показано далее, существует ряд классов графических

компонентов, языки которых отличаются, но между языками отсутствуют четкие границы, и они образуют систему языков с непрерывными переходами. Поэтому проблема определения языковой принадлежности самой компоненты и составляющих ее знаков заслуживает отдельного рассмотрения.

Решение вопросов многозначности и недетерминированности квантования графических и неоднородных компонентов на знаки и их сочетания, нормализации знаков, определения языковой принадлежности является основой для получения знаковых представлений научных документов. Эти вопросы будут рассмотрены во второй части этой статьи.

Если рассматривать все вербальные и невербальные компоненты научных документов некоторой электронной библиотеки, то для знакового представления каждого документа и всех его компонентов необходимо построить семиотическую систему для этой электронной библиотеки, включающей следующие системы знаков:

- традиционные вербальные системы знаков естественных языков (слова, устойчивые словосочетания и предложения),
- системы структурных знаков,
- системы графических знаков,
- системы неоднородных знаков (вербально-структурных, вербально-графических, структурно-графических и вербально-структурно-графических).

Этот перечень систем знаков соответствует предложенной типологии коммуникативных компонентов научных документов. В него, кроме систем знаков однородных компонентов, включены системы неоднородных знаков. Необходимость такого включения в семиотическую систему электронной библиотеки иллюстрирует рис. 2. На этом рисунке концепт "Течение реки, впадающей в озеро", графическая форма которого позволяет увидеть также и характер течения реки, может быть передан неоднородными вербально-графическими знаками или сочетанием взаимосвязанных графических и вербальных знаков. С помощью использования неоднородных знаков для этого концепта можно исключить этап описания взаимосвязей графических и вербальных знаков. Отметим также, что между разными сочетаниями вербально-графических знаков, отображающих один и тот же концепт, могут быть установлены отношения синонимии. Например, для знакового представления направления течения реки могут использоваться слова и символ "стрелка" (рис. 3).

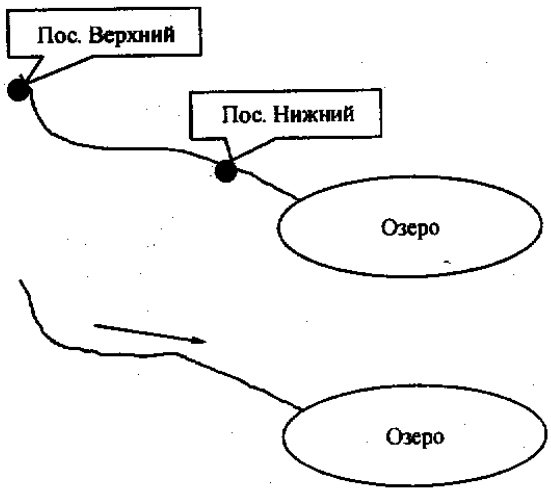

Рис. 3. Синонимичные формы представления направления течения реки [21]

Далее, говоря о принципах построения семиотических систем электронных библиотек, основное внимание будет уделено системам графических знаков. Вербальные и структурные компоненты, а также их знаковые системы будут рассматриваться только с точки зрения их сопоставления с графическими компонентами. Основная цель этого сопоставления заключается в том, чтобы составить перечень семиотических характеристик, единый для всех однородных компонентов, и определить допустимые значения семиотических характеристик для каждого вида компонентов научных документов.

4. СОДЕРЖАТЕЛЬНЫЕ АСПЕКТЫ ДОКУМЕНТОВ И СЕМАНТИЧЕСКОЕ ПРОСТРАНСТВО ЭЛЕКТРОННОЙ БИБЛИОТЕКИ

Структурные и семантические связи компонентов внутри научных документов задаются их авторами. Эти связи частично отражаются с помощью схем в электронных формах документов на первой стадии логико-семантического моделирования. Однако при интеграции всего корпуса документов и формировании на его основе электронных библиотек возникают новые семантические связи, включая и междокументальные. Одним из источников новых связей является тезаурус электронной библиотеки. Возможности тезауруса, с точки зрения формирования новых семантических связей, во многом зависят от концепции его построения и тех областей вербальных и невербальных сфер представления знаний, которые он охватывает.

Пространство знаний, отраженных и в корпусе научных документов, и в электронной библиотеке с учетом новых семантических связей и тезаурусных отношений, будем называть *семантическим пространством. Пространством семантического поиска* будем называть структурированное на области поиска семантическое пространство с установленными критериями упорядоченности и мерой близости между объектами поиска.

Цель второй стадии логико-семантического моделирования заключается в том, чтобы сформировать семантическое пространство на основе содержательных аспектов научных документов. Так как результатом первой стадии построения модели каждого документа являются его адресуемые однородные и неоднородные компоненты, то семантическое пространство документов электронной библиотеки можно схематично представить в виде сочетания областей содержательных аспектов однородных и неоднородных компонентов документов. Ограничимся областями однородных компонентов, и изобразим их на рис. 4 в виде четырех прямоугольников, обозначенных буквами A, B, C и D:

A — структурированные метаданные документов, компонентов и их фрагментов, включая атрибуты и/или библиографические описания документов (название документа, авторы, даты, издательство, ключевые слова и т. д.);

B — вербальные компоненты всех документов библиотеки;

C — структурные компоненты;

D — графические компоненты.

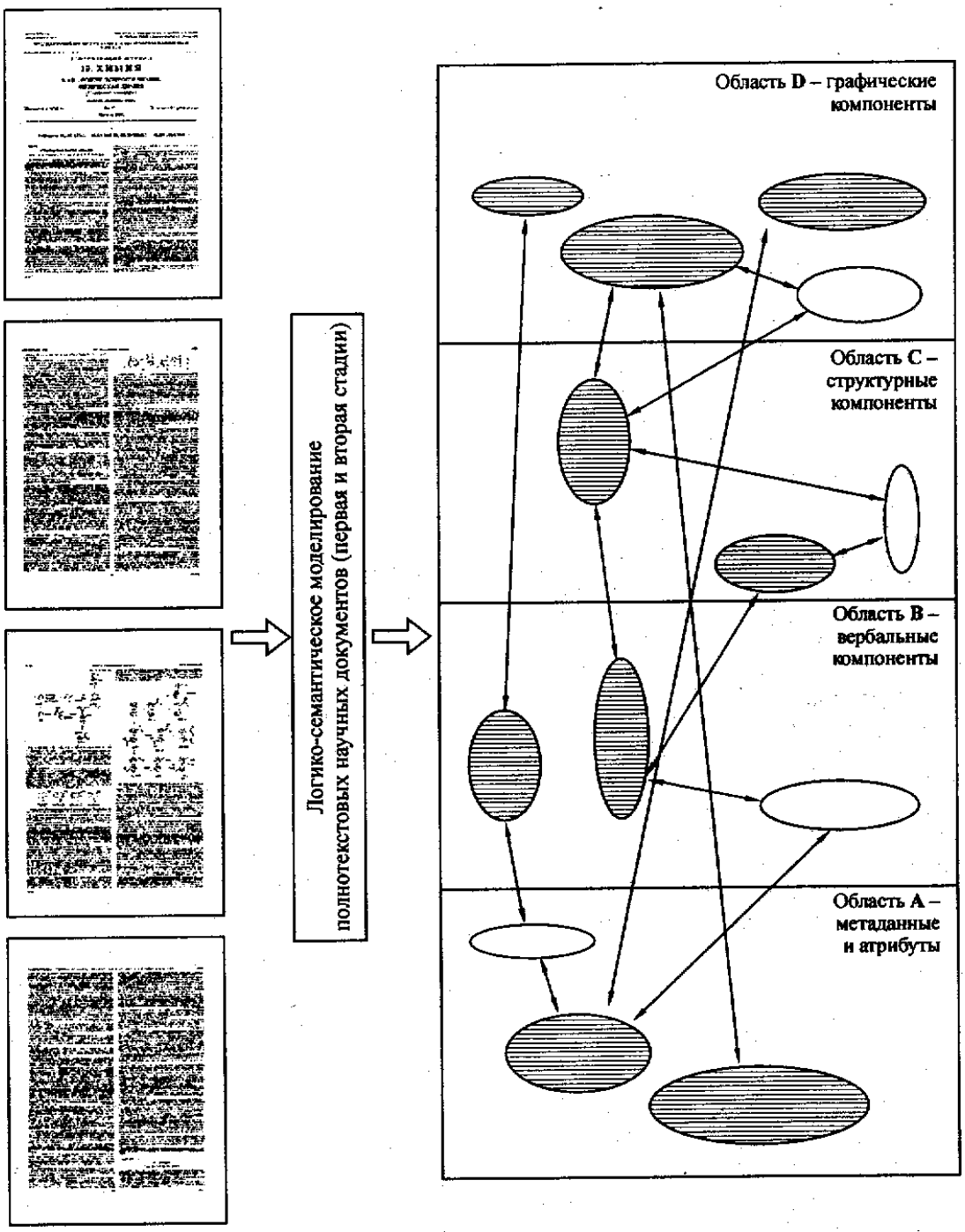

Рис. 4. Отображение содержательных аспектов документов в семантическое пространство электронной библиотеки

На рис. 4 не показаны четыре области неоднородных компонентов: вербально-структурные, вербально-графические, структурно-графические и вербально-структурно-графические. Все однородные компоненты (вербальные, структурные и графические) всех документов электронной библиотеки логически объединяются, соответственно, в областях B, C и D семантического пространства, а все метаданные — в области A.

Слева на рис. 4 изображены страницы моделируемых традиционных полнотекстовых научных документов на примере реферативного журнала ВИНИТИ. Концепты, которые представлены в исходных документах до установления междокументальных связей и использования тезауруса, схематично обозначены с помощью заштрихованных овалов.

Белые незаштрихованные овалы в областях семантического пространства на рис. 4 обозначают новые концепты, которые отсутствовали в исходных документах и были получены за счет использования тезауруса при интеграции документов и формировании семантического пространств. Наличие белых овалов во всех областях семантического пространства подразумевает создание тезауруса электронной библиотеки, который охватывает все три сферы представления знаний.

Разработка концепции создания такого тезауруса, который предлагается называть вербально-образным, является актуальной и нерешенной проблемой. Для всех трех сфер представления знаний необходимо иметь тезаурус, дескрипторы и знаки которого принадлежат всему спектру вербальных

и невербальных языков научных документов электронной библиотеки. Кроме того, для использования тезауруса при логико-семантическом моделировании документов с авторскими и слабоопределенными знаками необходимо предварительно их доопределить, т. е. соотнести форму каждого такого знака с его значением [17, 22, 23].

Представление в знаковой форме континуальных изображений в графических компонентах документов основывается на семиотической аппроксимации, которая рассматривается во второй части статьи, и решении задачи доопределения значений знаков с помощью вербально-образного тезауруса. Суть предлагаемого решения заключается в дискретизации континуума точек графических компонентов документов с помощью графических знаков и их сочетаний, заранее построенных на основе дескрипторов тезауруса.

За последние тридцать лет накоплен большой опыт проектирования и построения вербальных тезаурусов на основе знаков естественных языков [24–27]. Эти тезаурусы используются и при организации вербального поиска информации в электронных библиотеках, и для поддержки взаимодействия пользователей с многоколлекционными электронными библиотеками [28, 29].

В настоящее время имеется опыт построения тезаурусов и для графических материалов и изображений [30, 31]. Эти тезаурусы содержат вербальные дескрипторы и алфавитно-цифровые коды, соответствующие цветовой гамме и/или текстуре изображений. Однако эти тезаурусы не имеют образных дескрипторов и не охватывают ту сферу знаний, которые достаточно адекватно могут быть представлены только в невербальной форме.

5. СТРУКТУРА СЕМАНТИЧЕСКОГО ПРОСТРАНСТВА ЭЛЕКТРОННОЙ БИБЛИОТЕКИ

Семантическое пространство на рис. 4 схематично было изображено в виде объединения четырех областей A, B, C и D. Структурированность каждой области поиска не показана. Основываясь на предлагаемой типологии компонентов и принципах декомпозиции научных документов, можно более детально описать структуры областей и значения, принимаемые семиотическими характеристиками компонентов, объединяемых в рамках этих областей.

Структуры областей и значения семиотических характеристик иллюстрируются на примере политематического запроса, состоящего одновременно из четырех подзапросов. Этот составной запрос на поиск геохимических документов по содержательным аспектам компонентов, которые распределены по разным областям семантического пространства электронной библиотеки, схематично изображенного на рис. 5.

Первый подзапрос задает поиск сочетания геологических и топографических картообъектов, изображенных в документах электронной библиотеки, по их метаданным, т. е. в области A. Второй подзапрос — по лексике в компонентах тех же документов, включая текст разделов документов, подрисуночные подписи, а также текст из ячеек таблиц и диаграмм, т. е. в области B.

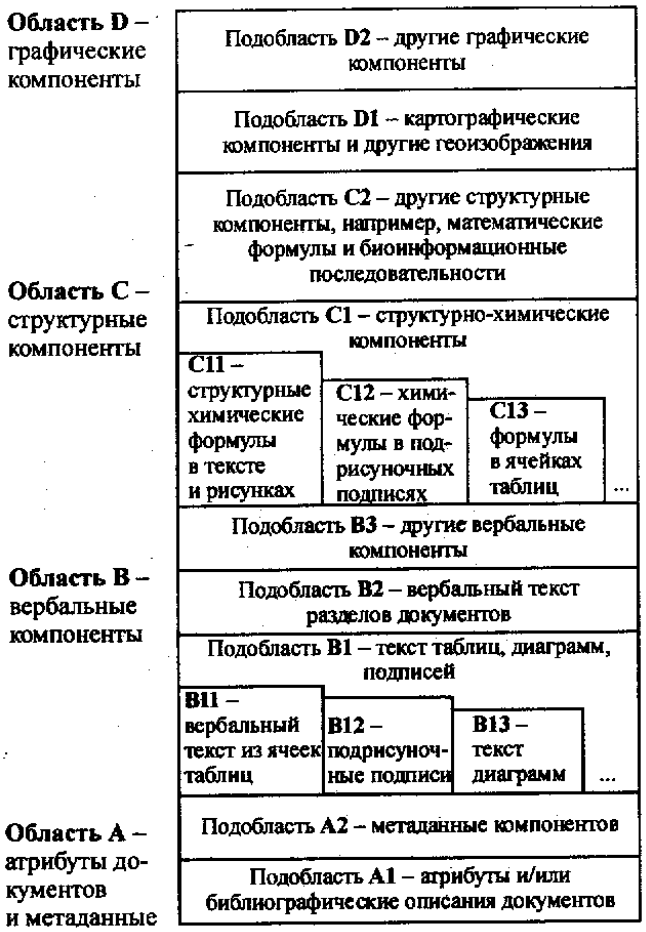

Рис. 5. Пример структуризации областей семантического пространства электронной библиотеки

Третий подзапрос задает поиск заданного фрагмента структурной химической формулы из тех же документов, включая формулы в вербальном тексте и рисунках, на полях графиков и рисунков, а также в ячейках таблиц, т. е. в области C. Четвертый подзапрос — по сочетаниям фрагментов изолиний геологических и топографических картообъектов в картах из тех же документов электронной библиотеки, т. е. в области D.

В соответствии с первым поисковым подзапросом выделим из области A две области второго уровня:

- атрибуты и библиографические описания документов — A1;
- метаданные графических компонентов и их смысловых фрагментов (в этом запросе — картообъектов) — A2.

Библиографический поиск и поиск по метаданным широко используются в традиционных структурированных базах данных и электронных библиотеках. Это достаточно известная и практически решенная задача [32, 33].

В соответствии с вторым поисковым подзапросом выделим из области B три области второго уровня:

- вербальный текст из ячеек таблиц, диаграмм и подрисуночных подписей — B1;
- вербальный текст разделов документов — B2;
- другие вербальные компоненты документов — B3.

Аналогичным образом в подобласти B1 можно выделить области третьего уровня: для текста из ячеек таблиц — B11, подрисуночных подписей — B12, текста диаграмм — B13. Для многоязычных электронных библиотек области третьего уровня можно разделить, в свою очередь, на области четвертого уровня по числу используемых естественных языков в области B, которые на рис. 5 не показаны.

Реализация второго подзапроса в электронных библиотеках полнотекстовых документов является во многом уже решенной задачей при достаточном уровне структурированности, разметки документов и их вербальных компонентов. Например, в проекте Иллинойского университета обеспечивается индексирование и полнотекстовый поиск документов, включая подрисуночные подписи [2]. Есть пример включения в область полнотекстового поиска вербальных компонентов, содержащихся в ячейках таблиц [34]. Таким образом, уже имеются примеры реализации полнотекстового поиска документов в структурированных областях вербальных компонентов.

В соответствии с третьим поисковым подзапросом из области C выделим две подобласти структурных компонентов:

- структурные химические компоненты — C1;
- математические формулы, биоинформационные последовательности и другие структурные компоненты — C2.

В свою очередь подобласть структурных химических компонентов также разделим на три области третьего уровня: структурные химические формулы в вербальном тексте разделов документов и рисунках, на полях графиков — C11, структурные химические формулы в подрисуночных подписях — C12, структурные химические формулы в ячейках таблиц — C13.

Реализация третьего подзапроса на поиск по фрагменту структурной химической формулы или реакции является уже решенной задачей в структурных химических базах данных [35]. Таким образом, при достаточной степени структурированности и разметки документов и их компонентов области третьего уровня C11, C12 и C13 также можно отнести к уже освоенным областям поиска.

В соответствии с четвертым поисковым подзапросом выделим из области D:

- картографические и другие геоизображения — D_1;
- отделив их от остальных графических компонентов — D2.

Для нашего примера подзапроса на поиск сочетания геологических и топографических картобъектов по заданным изолиниям необходимо рассмотреть только подобласть D1. В настоящее время в политематических электронных библиотеках научных документов отсутствуют службы поиска, которые позволяли бы реализовывать четвертый подзапрос поиска по контурам изолиний в иллюстрациях документов.

В электронных библиотеках для картографических и других геоизображений практически решены задачи поиска только по метаданным карт, карт-объектов и других геоиллюстраций [36]. Следовательно, поиск организован по алфавитно-цифровым (литерным) полям метаданных в области A семантического пространства электронной библиотеки, а не по содержательным аспектам и пространственным признакам графических компонентов в области D.

6. СЕМИОТИЧЕСКИЕ ХАРАКТЕРИСТИКИ КОМПОНЕНТОВ НАУЧНЫХ ДОКУМЕНТОВ

Рассмотрев пример структурированности областей семантического пространства, остановимся на отличительных признаках областей однородных компонентов, с точки зрения значений их семиотических характеристик.

По определению, знаки вербальных компонентов научных документов являются детерминированными, двумерными и статичными. Сочетания знаков являются линейно упорядоченными, а все знаки в сочетаниях сохраняют свойство дискретной отделимости одного от другого.

Знаки структурных компонентов являются детерминированными, одномерными или двухмерными, а сочетания знаков — дискретно отделимыми, полилинейно упорядоченными, или порядок и связи между литерами и знаками задаются с помощью сетевых, иерархических, реляционных и дискретных параметрических схем. Примером полилинейной упорядоченности являются последовательности аминокислот с одновременным указанием в параллельной строке их позиций в последовательности. В обобщенных документах могут встречаться динамичные и/или многомерные структурные компоненты, которые, как правило, имеют растровую или векторную и, редко, знаковую основу их построения.

Детерминированность очертаний знаков и дискретность их сочетаний в вербальных и структурных компонентах являются основными отличительными признаками областей B и C от области D. В знаковом представлении картографических и других геоизображений могут использоваться детерминированные, размытые, вероятностные и неопределенные по своим формам и очертаниям знаки, одномерные, двухмерные и статичные, а в обобщенных документах — динамичные и многомерные [9, 37]. Их сочетания могут быть дискретными, континуальными и дискретно-континуальными, знаки могут объединяться с помощью непрерывных многопараметрических, в том числе, пространственно-временных схем. При этом континуальные сочетания знаков в картографических изображениях редко однозначно и детерминированно разделяется на составляющие.

Например, в картографическом изображении холма с одной вершиной знак "вершина холма" и знаки "сектор склона холма" нельзя выделить однозначно и детерминировано, так как в знаковом представлении изображение холма может быть представлено в виде сочетания знака "вершина холма" и знаков "сектор склона холма" бесчисленным числом вариантов. В каждом варианте формы знаков могут отличаться, если эти отличия не влияют на содержательные аспекты знаков "вершина холма" и "сектор склона холма". Каждый

новый вариант можно получить за счет коррекции формы знаков без изменения их значений, т. е. достаточно в континуальном сочетании графических знаков, рассматриваемого как континуум точек, немного изменить формы этих знаков. При этом не должны затрагиваться те ключевые графические элементы знаков, которые отражают их содержательные аспекты и характерные пространственные признаки. Каждое такое изменение будет давать новый вариант знакового представления изображения холма в виде сочетания составляющих его графических знаков.

С одной стороны, графические компоненты, в общем случае, и картографические изображения, в частности, являются континуумами точек, из которых не всегда могут быть выделены знаки так, как это традиционно делается при выделении слов из вербальных выражений. С другой стороны, возможны варианты разделения континуума на знаки. При этом разные варианты объединяются едиными содержательными аспектами и знаков, и их сочетаний. Более того, возможен случай нечетких по форме знаков, когда можно наблюдать в графической компоненте размытые знаки с вполне определенными значениями, но невозможно выделить их однозначно или многовариантно по некоторым детерминированным и четки формам и границам.

Именно поэтому графические компоненты обладают такими значениями семиотических характеристик, которые отсутствуют в вербальных и структурных компонентах. Рассмотренные значения семиотических характеристик вербальных и невербальных компонентов научных документов предлагается сгруппировать по следующим пяти позициям:

- детерминированность/размытость очертаний знаков;
- статика/динамика форм знаков;
- одномерность/двухмерность/многомерность форм знаков;
- упорядоченность/неупорядоченность сочетаний знаков;
- дискретность/континуальность сочетаний знаков.

В предлагаемой типологии семиотических характеристик первая характеристика может принимать следующие значения (возможные значения остальных характеристик следуют из их названий) [9]:

- однозначность выделения в компоненте знаков с детерминированными формами;
- многовариантность выделения знаков с детерминированными формами;
- наблюдение в компоненте хотя бы одного нечеткого (размытого по очертаниям) знака;
- присутствие в компоненте знака с неопределенной формой.

Компоненты, принимающие первые два значения этой семиотической характеристики, при статичности, двумерности компонента и упорядоченности в нем сочетаний знаков могут быть представлены, как правило, в знаковой форме с использованием традиционных языков семантической разметки документов. Для остальных случаев потребуется разработка новых или существенная доработка существующих языков семантической разметки на вербально-образной семиотической основе.

Отметим, что разработка новых языков семантической разметки для графических компонентов зависит от нормализации графических знаков. При этом решение проблемы нормализации в общем случае должно учитывать все возможные значения первой семиотической характеристики, включая многовариантность выделения знаков, наличие нечетких и неопределенных по форме знаков и их сочетаний.

Для графических компонентов научных документов не всегда просто установить их языковую принадлежность. При этом четкая разделимость языков характерна только для вербальных и структурных компонентов. Для любой однородной структурной компоненты, относящейся к области второго уровня, используется, по определению, единственный язык, например, язык химических формул для C1 (см. рис. 5). В области B для любой моноязычной вербальной компоненты, относящейся к области четвертого уровня, используется, по определению, единственный естественный язык, например, русский или английский.

Таким образом, можно говорить о дискретности ряда языков для областей B и C. В подобласти D1 картографические иллюстрации и другие геоизображения представляют собой целый ряд классов графических компонентов, языки которых отличаются, но между языками отсутствуют четкие границы, и они образуют систему языков с непрерывными переходами. Иногда, в соответствии с непрерывной языковой системой, могут быть упорядочены и классы графических компонентов (см. рис. 10 в [21 с. 51]).

На этом рисунке в книге "Язык карты" в центральной его части в классах графических компонентов, обозначенных как "д", "е" и "ж", помещаются все известные карты, подробно характеризующие и пространственную, и содержательную определенности объектов, входящих в состав этих карт. Слева от центральной части располагаются изображения сеток картографических проекций с некоторыми элементами земной ситуации, отражающей контуры материков. Справа от центральной части располагаются картоподобные диаграммы. Еще правее находятся графики и диаграммы с указанием географической приуроченности [21, 38].

В гамме классов графических компонентов широкий спектр значений первой семиотической характеристики сочетается с непрерывностью языковой системы, что является фундаментальной проблемой неопределенности границ между языками этих классов, а часто и нечеткости языковой принадлежности знаков.

Отметим, что в семиотике эта проблема была сформулирована применительно к семиосфере культуры [39], но она является актуальной и для семантического пространства электронной библиотеки, если ставить задачу знакового представления на вербально-образной основе и организации семантического поиска во всем ее пространстве, а не только для метаданных, вербальных и структурных компонентов.

7. ЗАКЛЮЧЕНИЕ

Предложенный перечень семиотических характеристик коммуникативных компонентов научных документов и типология, принимаемых ими значений, существенно расширяют понятие "знак" в семиотике, в первую очередь, за счет возможной недетерминированности форм и очертаний знаков. Это

расширение понятия "знак" необходимо для получения знаковых представлений электронных форм вербальной и невербальной научной информации в процессе ее логико-семантического моделирования, а также для решения проблемы семантического поиска в библиотеках научных документов.

Именно появление феномена электронных библиотек стимулировало расширения понятия "знак". Для традиционных лингвистических, психологических и логических исследований характерно рассматривать формы, функции и значения знаков в ситуациях общения двух или большого числа индивидов. Более общий подход к исследованию знаковых систем, предложенный Г. П. Щедровицким и не связанный с описанием психических процессов и сознания индивидов, был основан на теории деятельности индивидов с различением типов деятельности. Ключевым положением этого подхода является механизм передачи в социуме знаковых образований, в том числе в виде научных документов [40].

Формирование электронных библиотек коренным образом изменяет механизм передачи, так как в социальных коммуникациях существенно возрастает роль поиска необходимых сведений, представленных в виде традиционных или обобщенных документов. А этот поиск выполняется компьютерными системами, возможности которых во многом зависят от спектра используемых в них систем знаков. Как появление в свое время телевидения привело к формированию новых видов знаковых систем [41], так и появление феномена электронных библиотек привело к необходимости переосмысления семиотических основ информатики и построения принципиально новых семиотических систем, необходимых для решения проблемы семантического поиска в электронных библиотеках.

СПИСОК ЛИТЕРАТУРЫ

1. Schatz B., Cole T. W., Hardin J. B. et al. Federating Diverse Collection of Scientific Literature // Computer.— 1996.— Vol. 29, № 5.— P. 28–36.

2. Schatz B., Mischo W., Cole T., Bishop A. et al. Federated Search of Scientific Literature // Computer.— 1999.— Vol. 32, № 2.— P. 51–59.

3. Wilensky R. Toward Work-Centered Digital Information Services // Computer.— 1996.— Vol. 29, № 5.— P. 37–44.

4. Gupta A., Santini S., Jain R. In Search of Information in Visual Media // Communications of the ACM.— 1997.— Vol. 40, № 12.— P. 35–42.

5. Lemke J. L. Multiplying Meaning: Visual and Verbal Semiotics in Scientific Text // J. R. Martin and R. Veel (Eds.) Reading science: Critical and functional perspectives on discourse of science.— London: Routledge, 1998.— P. 87–113.

6. Miller T. Visual Persuasion: A comparison of visuals in Academic Texts and the Popular Press // English for Specific Purposes.— 1998.— Vol. 17, № 1.— P. 29–46.

7. Johns A. M. The Visual and The Verbal: A Case Study in Macroeconomics // English for Specific Purposes.— 1998.— Vol. 17, № 2.— P. 183–197.

8. Зацман И. М. Электронные библиотеки научных документов в Интернет: структуризация, формальное описание и поиск невербальной информации // НТИ. Сер. 2.— 1998.— № 11.

9. Зацман И. М. Семантическое кодирование и разметка геолого-географических документов в политематических электронных библиотеках // Информационные технологии.— 2000.— № 11.— С. 2–11.

10. Eco U. A Theory of Semiotics.— Bloomington: Indiana University Press, 1976.— 356 p.

11. ГОСТ Р 51353-99 "Геоинформационное картографирование. Метаданные электронных карт. Состав и содержание", 1999.

12. Federal Geographic Data Committee, 1998. Content Standard for Digital Geospatial Metadata (**URL**: http://www.fgdc.gov/metadata/contstan.html).

13. Зацман И. М. Логико-семантические модели полнотекстовых научных документов // НТИ. Сер. 2.— 1999.— № 5.

14. Шемакин Ю. И., Романов А. А. Компьютерная семантика. — М.: НОЦ "Школа Китайгородской", 1995.— 344 с.

15. Степанов Ю. С. В мире семиотики // Семиотика: Антология / Сост. Ю. С. Степанов.— 2-е изд., испр. и доп.— М.: Академический проект, 2001.— С. 5–42.

16. Андрющенко В. М. Об организации архива источников Машинного фонда русского языка, их разметке и комментировании // Бюллетень Машинного фонда русского языка.— 1992.— № 1.— С. 3–44.

17. Zatsman I. M. Semantic Encoding and Markup of Georeferenced Documents in Polythematic Digital Libraries of Scientific Literature // Third All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections", Petrozavodsk, 11–13 September, 2001.— Petrozavodsk, 2001.

18. Lee K. H., Choy Y. C., Cho S. B. Geometric Structure Analysis of Document Images: A Knowledge-Based Approach // IEEE Transactions on Pattern Analysis and Machine Intelligence.— 2000.— Vol. 22, № 11.— P. 1224–1239.

19. Зацман И. М. Структуризация и семантическая разметка документов в электронных библиотеках // Тр. Междунар. семинара Диалог-99 по компьютерной лингвистике и ее приложениям. Т. 2.— Таруса, 1999.— С. 67–73.

20. Polzonetti G., Garravetta V., Russo M. V. et al. Phenylacetylene chemisorbed on Pt (111), reactivity and molecular orientation as probed by NEXAFS. Comparison with condensed multilayer and polyphenylacetylene // J. of Electron Spectroscopy.— 1999.— Vol. 98-99.— P. 175–187.

21. Лютый А. А. Язык карты: сущность, система, функция.— М.: ИГ АН СССР, 1988.

22. Барт Р. Основы семиологии // Французская семиотика: От структурализма к постструктурализму.— М.: изд. группа "Прогресс", 2000.— С. 247–310.

23. Соломоник А. Семиотика и лингвистика.— М.: "Молодая гвардия", 1995.— 352 с.

24. Шемакин Ю. И. Тезаурус в автоматизированных системах управления и обработки информации.— М.: Воениздат, 1974.— 192 с.

25. Тезаурус научно-технических терминов / Под. ред. Ю. И. Шемакина.— М.: Воениздат.— 1972.— 672 с.

26. Лукашевич Н. В. От общеполитического тезауруса к тезаурусу русского языка в контексте автоматической обработки больших массивов текстов // Тр. Междунар. семинара Диалог-99 по компьютерной лингвистике и ее приложениям. Т. 2.— Таруса, 1999.— С. 184–190.

27. Лукашевич Н. В., Добров Б. В. Исследование тематической структуры текста на основе большого лингвистического ресурса // Тр. Междунар. семинара Диалог-2000 по компьютерной лингвистике и ее приложениям. Т. 2.— Протвино, 2000.— С. 252–258.

28. Казаков Е. Н. Формирование и ведение тезауруса в составе посредника между пользователями и сетью электронных библиотек // Тр. Первой Всерос. науч. конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции".— Спб., 1999.— С. 85–88.

29. Казаков Е. Н., Сомин Н. В. Тестирование соответствия тезауруса посредника лексике отдельных коллекций // Тр. Второй Всерос. науч. конф. "Электронные библиотеки: перспективные методы и технологии, электронные коллекции".— Протвино, 2000.— С. 234–237.

30. Library of Congress' Thesaurus for Graphic Material (TGM) (URL: http://lcweb.loc.gov/rr/print/tgm1/toc.html).

31. Manjunath B. S. An Image Thesaurus for Content Based Search Using Texture and Color (URL: http://www.cs.pitt.edu/~panos/idm98/Imported/manj.html).

32. Черный А. И. Введение в теорию информационного поиска.— М.: Наука, 1975.— 240 с.

33. Солтон Дж. Динамические библиотечно-информационные системы.— М.: Мир, 1979.

34. Croft W. B. NSF Center for Intelligent Information Retrieval // Comm. of the ACM.— 1995.— Vol. 38, № 4.— P. 42–43.

35. Авакян В. Г., Трепалин С. В., Воронежева Н. И., Чуракова Н. И. Поиск реакций по изменяющимся фрагментам в программе графической обработки структурной химической информации CBASE // Материалы 3-й Междунар. конф. НТИ-97 "Информационные ресурсы. Интеграция. Технологии", Москва, 26–28 ноября 1997 г.— М.: ВИНИТИ, 1997.— С. 9–12.

36. Зацман И. М., Лютый А. А., Мартыненко А. И. Семантический поиск в электронных геобиблиотеках // Системы и средства информатики. Вып. 10.— М.: Наука, 2000.— С. 192–204.

37. Зацман И. М. Семантическое поле поиска геодокументов в политематических электронных библиотеках // Тр. Междунар. семинара Диалог-2000 по компьютерной лингвистике и ее приложениям. Т. 2.— Протвино, 2000.— С. 148–158.

38. Лютый А. А. Система "язык карты", основные черты устройства // Теоретические аспекты географии. Вопросы географии.— М.: Мысль, 1984.— Вып. 122.— С. 40–56.

39. Лотман Ю. М. Внутри мыслящих миров. Человек — текст — семиосфера — история.— М.: "Языки русской культуры", 1996.— 464 с.

40. Щедровицкий Г. П. О методе семиотического исследования знаковых систем // Семиотика и восточные языки.— М.: Наука, 1967.— С. 19–47.

41. Аронсон О. В. Технологии сообщества // Традиционная и современная технологии.— М.: ИФРАН, 1999.— С. 126–127.

Материал поступил в редакцию 12.02.01