КОНЦЕПТУАЛЬНЫЙ ПОИСК ИНФОРМАЦИОННЫХ ОБЪЕКТОВ В ЭЛЕКТРОННЫХ БИБЛИОТЕКАХ НАУЧНЫХ ДОКУМЕНТОВ[1)]

И.М. Зацман,
Институт проблем информатики РАН,
e-mail: igor@170.ipi.ac.ru

В схеме концептуального поиска предлагается выделить два этапа отображения: «знания – информация» и «информация – цифровые коды». Информация рассматривается как сочетание вербальных и невербальных знаковых форм представления знания в электронных библиотеках научных документов, созданных человеком, а цифровые коды – это последовательности цифр «0» и «1», предназначенные для их обработки компьютерными системами. Основное содержание доклада заключается в структуризации проблемы концептуального поиска на задачи с использованием схемы двухэтапного отображения в компьютерном представлении знаний.

1. Введение

Авторы хрестоматийных работ сороколетней давности, исследуя вопросы реализации человеко-машинного интерфейса, различали соответствие слов и соответствие концептов [i, ii]. Концептуальный поиск, то есть поиск на основе соответствия концептов, является в настоящее время актуальной проблемой для электронных библиотек с управляемыми информационными ресурсами и для WWW. Литературы, посвященной проблеме концептуального поиска с учетом невербальных форм представления научных знаний почти нет – в основном эта проблема рассматривается применительно к текстовой информации на естественных языках, то есть рассматриваются, как правило, *вербальные концепты* [iii].

Проблема поиска на основе соответствия концептов в электронных библиотеках с учетом невербальных форм представления научных знаний включает ряд задач, формулировка которых является основной целью доклада. С одной стороны, она является одной из тех фундаментальных проблем информатики, решение которых необходимо для создания электронных библиотек новых поколений. С другой стороны, это междисциплинарная проблема, так как для ее постановки и решения, кроме теоретических основ информатики, необходимо привлечь и адаптировать результаты, полученные в когнитологии, психологии, семиотике и прикладной лингвистике.

2. Когнитивная схема концептуального поиска

Основная цель настоящего параграфа заключается в описании предлагаемой схемы концептуального поиска. В описании используется словосочетание «*мультимодальное представление знаний*». Его использование предполагает следующую трактовку понятия «представление знаний» - в информационных системах знания человека могут быть представлены в виде вербальных и невербальных информационных объектов. Смысл словосочетания «мультимодальное представление знаний» поясним на примере.

Предположим, что некоторый концепт можно выразить словом (например, словом «любить») и пиктограммой (например, в виде сердечка). Подобные концепты, которые имеют несколько информационных модальностей

[1)] Работа выполнена при частичной поддержке РФФИ в рамках проекта 01-06-80332

представления (для приведенного примера их указано две, вербальная и образная модальности), будем называть *полимодальными* или онтологическими двойниками[2)].

Если в электронной библиотеке, поддерживающей две модальности – вербальную и образную – и в запросе на концептуальный поиск пользователь использовал пиктограммы, то поисковая система, учитывающая свойство мультимодального представления знаний, должна искать, кроме информационных объектов с пиктограммами, и те объекты, в которых **этот же концепт** выражен словами.

Таким образом, при постановке проблемы концептуального поиска (ПКП) в электронных библиотеках, поддерживающих свойство мультимодального представления знаний, необходимо учитывать, что искомые сведения, которые пользователи рассчитывали найти в информационной системе в виде сочетания слов, могут быть представлены в иных модальностях, а не только в вербальной. Поэтому, в постановке ПКП, кроме вербальной синонимии, предлагается учитывать онтологических двойников и весь спектр информационных модальностей форм их представления, поддерживаемых в электронной библиотеке.

Рассмотрение предлагаемой схемы концептуального поиска начнем со списка составляющих эту схему процессов с использованием рис. 1. На этом рисунке условно обозначены восемь следующих процессов:

1. представление авторами научных документов своих *знаний* в виде вербальных *информационных* объектов (например, слова, словосочетания, предложения, текстовые параграфы) и невербальных *информационных* объектов (например, формулы, таблицы, диаграммы, графики, карты, схемы или рисунки);
2. декомпозиция научных документов и *цифровое кодирование* вербальных и невербальных *информационных* объектов, входящих в их состав, и отношений между объектами;
3. организация хранения и индексирование *цифровой* информации загруженных документов с использованием вербальных и невербальных элементарных единиц представления *знаний* и схем их упорядочивания, например, в виде мультимодального тезауруса электронной библиотеки;
4. представление пользователем электронной библиотеки искомых *концептов* в виде сочетаний вербальных и невербальных *информационных* объектов, которые будем называть поисковыми запросами;
5. декомпозиция поисковых запросов и кодирование их вербальных и невербальных информационных объектов в виде цифровой информации и отношений между объектами с использованием вербальных и невербальных элементарных единиц представления знаний и схем их упорядочивания;
6. *концептуальный* поиск по *цифровой* информации поисковых запросов на основе соответствия и/или близости элементарных единиц представления *знаний* и их сочетаний с использованием критериев близости, заданных пользователем, и извлечение найденной *цифровой* информации из электронной библиотеки;
7. восстановление средствами электронной библиотеки авторских вербальных и невербальных *информационных* объектов на основе найденной *цифровой* информации;
8. семантическая интерпретация **пользователем** восстановленных авторских *информационных* объектов (как он их понимает) и сравнение интерпретированных *концептов* (то есть результатов пользовательской интерпретации) с искомыми концептами, выраженными **им** в виде поискового запроса.

Цифрами от 1 до 8 на рис. 1 обозначены перечисленные в списке процессы. Схему, включающую эти процессы, предлагается назвать *когнитивной схемой концептуального поиска*. Отметим, что предлагаемая когнитивная схема охватывает и тот частный случай, когда автор и пользователь являются одним и тем же лицом.

На рис. 1 выделены три уровня когнитивной схемы (верхний, средний и нижний) в точном соответствии с триадой терминов «знания – информация – коды». Граница между верхним и средним уровнями условно обозначена пунктирной линией, между средним и нижним – штрих-пунктирной линией.

3. Знания – информация – коды

Справа на рис. 1 изображена триада терминов «знания – информация – коды», при использовании которой предполагается, что эти три термина не являются синонимами и выражают три разных концепта. Термин «знание» используется в книге только применительно к человеку – автору документов и пользователям электронной

[2)] Идея использования словосочетания ontological counterparts принадлежит анонимному рецензенту конференции ECDL2002, которую он высказал в отзыве на заявленный доклад автора

библиотеки. Термины «информация» и «информационные объекты» используются только применительно к вербальным и невербальным знаковым формам представления знаний в электронных библиотеках научных документов, созданных человеком. Термины «цифровые коды», «коды», «цифровая информация» и «цифровые объекты» в докладе рассматриваются как синонимы и используются только применительно к компьютерным цифровым кодам, точнее последовательностям цифр «0» и «1».

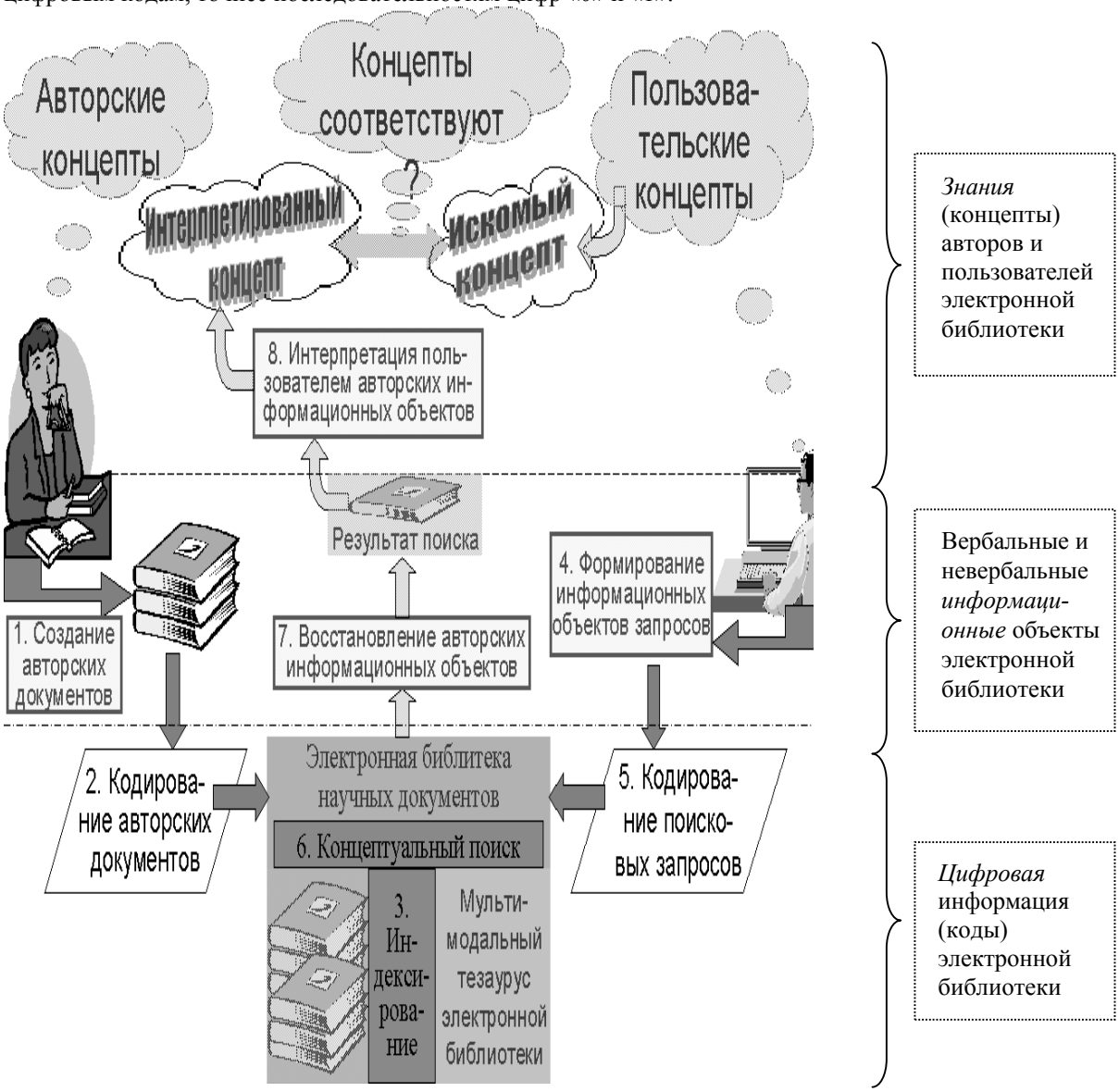

Рис. 1. *Когнитивная схема концептуального поиска*

Известны и другие подходы к толкованию понятий «информация (информационные объекты)», в которых не проводится четкое разграничение содержания понятий «информация» и «знания» [iv]. Предлагаемое в когнитивной схеме концептуального поиска разграничение содержания понятий «знание» и «информация» во многом соответствует положениям теории компьютерной семантики Ю.И. Шемакина и А.А. Романова. В компьютерной семантике информация рассматривается как результат отображения знаний о внешнем объекте. Информация приобретает смысл только при наличии интерпретатора, системы знаний субъекта в рамках его целеустремленной деятельности и при наличии потребностей в информации ([v], стр. 33).

В теории компьютерной семантики используется также термин «данные» для описания взаимодействия между компьютерной семантической системой и внешними объектами (внешними по отношению к системе). В этой теории понятие «цифровые коды» не выделяется и, соответственно, не проводится граница между информацией и кодами в компьютерной семантической системе.

Можно отметить два отличительных аспекта рассматриваемого варианта когнитивной схемы по сравнению с аналогичной схемой в книге Ю.И. Шемакина и А.А. Романова ([v], стр. 32). Во-первых, на рис. 1 не

рассматриваются внешние объекты и взаимодействие с ними учитывается только опосредованно через субъектов (авторов и пользователей). Во-вторых, в когнитивной схеме проводится разграничение содержания понятий «информация (информационные объекты)» и «цифровые коды».

Предлагаемое в когнитивной схеме концептуального поиска разграничение содержания понятий «информация» и «цифровые коды» имеет несколько общих позиций с описанием множественной модели данных, предложенной М.М. Гилулой [vi]. Он пишет: "…*информация кодируется с помощью данных и извлекается путем их декодирования и интерпретации*. При этом *коммуникационный процесс*, в котором участвует информационная система (ИС), можно упрощенно считать состоящим из следующих основных этапов:

1. Преобразование информация – данные. Исходная информация о предметной области кодируется с помощью данных для ввода в ИС, обработки и выдачи по запросам пользователей.
2. Преобразование данные – данные. Данные загружаются, обрабатываются и выдаются в ответ на запросы по реализованным в системе алгоритмам. Действия ИС на этом этапе не зависят от закодированной в данных информации.
3. Преобразование данные – информация. Выданные в ответ на запрос данные декодируются и интерпретируются пользователем с целью получения информации." ([vi], стр. 12).

При очевидной аналогии разграничения содержания понятий «информация» и «коды» в когнитивной схеме концептуального поиска и понятий «информация» и «данные» во множественной модели имеются и принципиальные отличия. При описании этой модели М.М. Гилула применяет термин «данные» не только к цифровым кодам компьютерных систем, но и к вербальным знаковым формам представления знаний (например, к библиотечным каталожным карточкам), что размывает границу между терминами «информация» и «данные» в его модели. В когнитивной схеме к вербальным знаковым формам представления знаний применяется только термин «информация», что позволяет провести более четкую границу между средним и нижним уровнями схемы.

И еще одно отличие, которое необходимо отметить, заключается в отсутствии во множественной модели невербальных форм представления знаний, так как набор типов объектов множественной модели по определению основан на сочетаниях и вложениях конечных цепочек литер ([vi], стр. 78). В когнитивной схеме набор типов объектов может включать изображения, структурные химические формулы и другие невербальные информационные объекты.

Подводя итоги, отметим, что перечисленные отличия не претендуют на полноту сравнения множественной модели данных и когнитивной схемы концептуального поиска.

Ряд авторов не выражает имеющееся у них разграничение содержания понятий «информация» и «цифровые коды» разными терминами. Например, D. McArthur в монографии «Информация, ее формы и функции» пытается разграничить эти два понятия и, говоря по сути о компьютерных цифровых кодах, использует словосочетание «информация в техническом смысле слова» ([vii], стр. 85). Однако во введении к этой монографии McArthur использует термин «информация» и для обозначения цифровых кодов на перфокартах, в калькуляторах и компьютерах ([vii], стр. ix), что размывает в его концепции форм и функций информации границу между терминами «информация» и «информация в техническом смысле слова». В когнитивной схеме в подобных случаях (перфокарты, калькуляторы и компьютеры) используется только термин «коды».

Необходимо сделать два предварительных замечания до описания процессов, изображенных на рис. 1. Во-первых, *далеко не всегда процессы приобретения, представления и передачи знаний фиксируются и документируются*. Возможны случаи, когда стабильных «информационных следов», скажем, в виде публикаций или отчетов, не остается (это может быть недокументированное устное консультирование, обмен персоналом, совещания и т.п. [viii, стр. 7].

В классификации технологий обработки/передачи знаний по модели Nonaka эти недокументированные случаи выделены как один из четырех основных видов технологий, который получил название “socialization process”, что на русский язык можно перевести как «процесс обобществления» [iv].

В рассматриваемом варианте когнитивной схемы концептуального поиска учитываются только те знания и концепты, которые зафиксированы и представлены в виде вербальных и/или невербальных информационных объектов как стабильных «информационных следов» процессов представления знаний.

Во-вторых, для постановки ПКП необходимо рассмотреть перечисленные восемь процессов и ответить на следующие вопросы:

1. Что понимается под вербальными и невербальными *элементарными единицами представления знаний* в электронных библиотеках?

2. Как соотносятся элементарными единицами представления знаний и информационные объекты?

В докладе нет ответов на эти вопросы и завершенных формулировок всех задач ПКП. Основная цель доклада – структуризация ПКП и постановка вопросов. В качестве примера, иллюстрирующего эти два вопроса отметим, что для текстовой информации на алфавитных естественных языках вербальными элементарными единицами представления знаний являются слова и устойчивые словосочетания.

4. Процессы когнитивной схемы концептуального поиска

В этом параграфе кратко рассмотрим шесть из восьми перечисленных процессов когнитивной схемы концептуального поиска. Процесс фиксации знаний в виде вербальных и невербальных *информационных* объектов, входящих в состав научных документов, упоминается в схеме на рис. 1 два раза: один раз для авторских и второй раз для пользовательских концептов (см., соответственно, 1 и 4 пункты в списке процессов).

Два раза в схеме упоминается процесс отображения вербальных и невербальных *информационных* объектов в цифровые коды: для научных документов и поисковых запросов (см., соответственно, 2 и 5 пункты списка).

Таким образом, авторские и пользовательские концепты (знания) сначала фиксируются в виде сочетаний вербальных и невербальных информационных объектов (**первый этап**), а затем этим объекты ставятся в соответствие цифровые коды (**второй этап**).

Два прямых последовательных перехода, сначала с верхнего уровня «знания (концепты)» на средний «информационные объекты», а затем со среднего уровня на нижний «коды» будем называть двухэтапным мультимодальным отображением знаний в цифровые коды электронной библиотеки. Два обратных последовательных перехода будем называть двухэтапной семантической интерпретацией информации, представленной в виде цифровых кодов электронной библиотеки.

Словосочетание «мультимодальное отображение» обозначает возможность использовать на среднем уровне схемы сочетания информационных объектов разных модальностей, то есть для фиксации своих концептов авторы и пользователи могут использовать и вербальные, и невербальные информационные объекты.

Таким образом, первый этап мультимодального отображения на рис. 1 изображен в виде переходов от верхнего уровня к среднему. На схеме прямые переходы первого этапа условно обозначены двумя прямоугольными: «1. Создание авторских документов» и «4. Формирование информационных объектов запросов». Эти два прямоугольника на рис. 1 изображены на среднем уровне, к которому относятся результаты прямых переходов первого этапа мультимодального отображения знаний.

Обратный переход со среднего уровня на верхний условно обозначен одним прямоугольником «8. Интерпретация пользователем авторских информационных объектов», то есть семантическая интерпретация на основе восстановленных информационных объектов, так как он их понимает. Этот прямоугольник на когнитивной схемы изображен на верхнем уровне, к которому относятся результаты процесса семантической интерпретации сочетаний вербальных и/или невербальных информационных объектов. В общем случае предполагается, что пользовательская интерпретация может не совпадать с исходным авторским концептом. Поэтому релевантность найденных в электронной библиотеке данных оценивается пользователем с точки зрения соответствия его искомого концепта и его интерпретации авторских информационных объектов.

Второй этап мультимодального отображения на рис. 1 изображен в виде переходов от среднего уровня к нижнему. На схеме прямые переходы второго этапа условно обозначены двумя параллелепипедами: «2. Кодирование авторских документов» (более точно, кодирование структуры авторских документов, их вербальных и невербальных информационных объектов) и «5. Кодирование поисковых запросов». Отметим, что эти параллелепипеды на когнитивной схеме условно изображены на нижнем уровне, к которому относятся результаты прямых переходов второго этапа мультимодальных отображений, а именно, цифровые коды электронной библиотеки.

Обратный переход с нижнего уровня на средний условно обозначен одним прямоугольником «7. Восстановление авторских информационных объектов» на основе найденных в электронной библиотеке цифровых кодов.

В заключение этого параграфа отметим один частный случай мультимодального представления знаний, когда из всех возможных невербальных форм в электронной библиотеке обрабатываются только изображения. Этот случай предлагается называть вербально-образным представлением знаний.

5. Задачи проблемы концептуального поиска

Состав первой задачи ПКП определим как последовательное объединение двух первых процессов когнитивной схемы. Аналогичным способом сформулируем еще две задачи ПКП, объединив четвертый и пятый, а также седьмой и восьмой процессы. Разница только в том, что для объединения четвертого и пятого процессов используются два прямых перехода: *знания → информация* и *информация → коды.* А для объединения седьмого и восьмого процессов используются два обратных перехода: *коды → информация* и *информация → знания*.

С учетом трех попарных объединений предлагается выделить следующие пять задач ПКП в электронных библиотеках:

1. формирования электронных (цифровых) документов, включая цифровое кодирование вербальных и невербальных информационных объектов, входящих в их состав, и контекстных отношений между объектами;
2. организация хранения и индексирование цифровой информации документов электронной библиотеки с использованием элементарных единиц представления знаний и схем упорядочивания элементарных единиц;
3. формирования поисковых запросов, включая цифровое кодирование вербальных и невербальных информационных объектов, входящих в их состав, и контекстных отношений между объектами;
4. концептуальный поиск по цифровой информации поисковых запросов на основе соответствия и/или близости элементарных единиц с использованием критериев близости, заданных пользователем, и извлечение найденной цифровой информации;
5. семантическая интерпретация пользователем восстановленных на основе найденной цифровой информации авторских информационных объектов и сравнение интерпретированных концептов (то есть результатов пользовательской интерпретации) с искомыми концептами, выраженными им в виде сочетания вербальных и невербальных информационных объектов поискового запроса.

6. Заключение

В соответствии с основной целью доклада определены границы пяти основных задач, но их формулировки содержат не определенное понятие «невербальные элементарные единицы представления знаний». Если для вербальных текстов на алфавитных естественных языках элементарные единицы представления знаний– это слова и словосочетания, то для многих невербальных информационных объектов это понятие является предметом актуальных исследований в семиотике.

Более того, для изображений к концу прошлого века стали доминировать те концепции в семиотике, в которых утверждается невозможность определения образных (визуальных) элементарных единиц представления знаний также, как определяются вербальные знаки, составляющие тексты на естественных языках [ix, x].

Приведенные формулировки задач не являются завершенными до тех пор, пока не будет найден подход к определению понятия «невербальные элементарные единицы представления знаний» с известной сферой их конвенциональности, что позволило бы называть их невербальными знаками.

Для изображений в работах [xi, xii] было предложено определить понятие «образные элементарные единицы представления знаний» через образные дескрипторы вербально-образного тезауруса электронной библиотеки. При этом значения образных дескрипторов определялись в рамках известных научных систем классификаций, что позволяло определить сферу конвенциональности образных дескрипторов и считать их образными знаками электронной библиотеки в рамках определенной сферы конвенциональности. Однако в этом случае завершенность формулировок задач будет иметь отношение только к вербальным текстам и изображениям, исключая все другие возможные невербальные информационные объекты.

Литература

i. Licklider J.C.R., Clark W.E. On-line man-computer communication // AFIPS Proceedings Spring Joint Computer Conference (May 1-3, 1962) - Vol. 21, 1962.- pp. 113-128.

ii. Licklider J.C.R. Libraries of the Future.– Cambridge: MIT Press, 1965.– 220 pp.

iii. Schatz B.R. Information Retrieval in Digital Libraries: Bringing Search to the Net // Science Magazine.- Vol. 275, No. 5298, 1997.- pp. 327-334.

iv. Marwick A.D. Knowledge management technology // IBM Systems Journal – Vol. 40, No. 4, 2001.– pp. 814-830.

v. Шемакин Ю.И., Романов А.А. Компьютерная семантика. – М.: НОЦ «Школа Китайгородской», 1995.- 344 стр.

vi. Гилула М.М. Множественная модель данных в информационных системах.– М.: Наука, 1992.- 208 с.

vii. McArthur D. Information, its forms and functions: The elements of semiology.- Lewinton: The Edwin Mellen Press, Ltd., 1997.– 228 pp.

viii. Инновационная система России: модель и перспективы ее развития. Вып. 1. Анализ мирового опыта формирования и функционирования инновационных систем в контексте развития российской национальной модели.– М:. Изд-во РУДН, 2002.- 84 с.

ix. Sonesson G. Die Semiotik des Bildes. Zum Forschungsstand am Anfang der 90er Jahre, in *Zeitschrift für Semiotik,* 15: 1—2, 1993; ss 131—164 (перевод на английский язык: http://www.arthist.lu.se/kultsem/sonesson/pict_sem_1.html).

x. Jorna R.J., Heusden B. Signs, search and communication: Towards an empirical future for semiotics. In: Jorna R.J., Heusden B., Posner R. (Eds.) Signs, search and communication: Semiotics aspects of artificial intelligence.- Berlin: Walter de Gruyter, 1993.- pp. 1-21.

xi. Зацман И.М. Визуально-мотивированное представление знаний в электронных библиотеках научных документов // Труды 4-й Всероссийской конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (Дубна, 15-17 октября 2002г.): В 2-х томах. Т. 1.- Дубна: ОИЯИ, 2002.- С. 120-135.

xii. Zatsman I. Pictorial Signs for Geoimages in Digital Libraries // 5th International Interdisciplinary Symposium "PICTURE LANGUAGE - VISUALIZATION - DIAGRAMMATICS", December 6-8, 2002, Vienna. Abstracts. – Vienna: Institute for Socio-Semiotic Studies, 2002.– p. 21.