УДК 004.62

ФОРМАЛИЗАЦИЯ ПРОЦЕССОВ ФОРМИРОВАНИЯ ПОЛЬЗОВАТЕЛЬСКИХ КОЛЛЕКЦИЙ В ЦИФРОВОМ ПРОСТРАНСТВЕ НАУЧНЫХ ЗНАНИЙ

Н. Е. Каленов, И. Н. Соболевская, А. Н. Сотников

Межведомственный Суперкомпьютерный Центр РАН (МСЦ РАН) – филиал Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», г. Москва

nkalenov@jscc.ru, ins@jscc.ru, ansotnikov@jscc.ru

Аннотация

Исследована задача формирования цифрового пространства научных знаний (ЦПНЗ). Рассмотрено отличие этого понятия от общего понятия пространства знаний. ЦПНЗ представлено как множество, содержащее объекты, верифицированные мировым научным сообществом. Формой структурированного представления цифрового пространства знаний является семантическая сеть, основной принцип организации которой основан на системе классификации объектов и последующем построении их иерархии, в частности, по принципу наследования. Введена классификация объектов, составляющих контент ЦПНЗ. Предложена модель ЦПНЗ как совокупности непересекающихся множеств, содержащих цифровые образы реальных объектов и их характеристики, обеспечивающие отбор и визуализацию объектов в соответствии с многоаспектными пользовательскими запросами. Определено понятие пользовательской коллекции, предложена иерархическая классификация типов пользовательских коллекций. Использование понятий теории множеств при построении ЦПНЗ позволяет разбивать информацию по уровням детализации и формализовать алгоритмы обработки пользовательских запросов, что проиллюстрировано конкретными примерами.

Ключевые слова*: семантическая сеть, информационное пространство, научные знания, электронная библиотека, уровни детализации, иерархия информационных объектов.*

ВВЕДЕНИЕ

Информация играет центральную роль во многих сферах нашей жизни. Развитие информационных и вычислительных технологий расширило возможности для сбора, анализа, распространения, обработки и использования научной информации.

Современные потребности в профессиональной информации требуют развития пространства знаний, представляющего собой цифровую среду, в которую интегрированы информационные ресурсы и сервисы из разных областей науки, культуры и образования. Частью общего пространства знаний является цифровое пространство научных знаний (ЦПНЗ), отличающееся от других составляющих общего пространства (в частности, такого, как Википедия) тем, что информационные объекты, представленные в ЦПНЗ, верифицированы мировым научным сообществом и отделены от информационных объектов, которые носят идеологический, религиозный и другой спорный с научной точки зрения характер [1].

Поток запросов в ЦПНЗ часто является непрерывным, быстро меняющимся во времени, не всегда предсказуемым и неограниченным по форме запроса. Программное обеспечение, обрабатывающее такие запросы, не может позволить себе хранить и «пересматривать» параметры запроса, часто требующего быстрого ответа в режиме реального времени. Требования точности поиска данных в ЦПНЗ (в отличие от общих поисковых машин интернета) обусловливают необходимость разработки специальных методов обработки поисковых запросов с обеспечением достаточно точного отображения текста запроса на пространство метаданных, описывающих те или иные объекты ЦПНЗ. Метаданные ЦПНЗ, в свою очередь, включают не только наборы ключевых слов, но и более сложные структуры, например, иерархические классификационные системы.

Формой структурированного представления цифрового пространства знаний является семантическая сеть, основной принцип организации которой основан на системе классификации объектов и последующем построении их иерархии, в частности, по принципу наследования: «макроэкономика» – раздел «экономики», «поэтический сборник» – издание и т. п. В соответствии с этим принципом объекты классифицируются на некоторое число категорий или классов на основании их общих свойств.

Большинство цифровых коллекций данных представляет собой разнородную информационную сеть, связывающую объекты различного типа. Например, электронная публикация (тип объекта – «книга»), помимо простого текста, содержит дополнительную информацию, такую, как автор публикации (тип объекта – «персона»), год издания, издательство, место издания и т. п. В свою очередь, объект «персона», кроме последовательности символов, задающих фамилию, связан с биографией, областью научных интересов («тематика объекта») и т. п. Таким образом, от объекта «публикация» может быть установлена связь с другим «объектом» («автор»), текстом этой публикации, «тематикой объекта» и т. п. В общем случае ЦПНЗ должно поддерживать различные типы связей между его элементами – как внутри одного класса объектов (в частности, рекурсивную иерархическую связь), так и между объектами различных классов.

Вопросу построения тематических иерархий, иерархии понятий, объектных моделей и т. д., обеспечивающих иерархическую организацию данных на разных уровнях детализации и имеющих такие приложения, как задачи веб-поиска и просмотра, посвящено значительное количество исследований [2,3, 4].

В [5] описан алгоритм NetClus, который позволяет устанавливать связи между многотипными объектами для создания высококачественных сетевых кластеров. Алгоритм NetClus позволяет переупорядочивать объекты атрибутов в каждом вновь определенном сетевом кластере.

В работе [6] нами предложена иерархия представления объектов в среде электронной библиотеки, которую можно рассматривать как прообраз ЦПНЗ.

В данной работе мы рассматриваем ЦПНЗ в аспекте теории множеств, что позволяет подойти к вопросам построения пространства и работы с ним с новой точки зрения.

1. СТРУКТУРА ЦПНЗ

Пусть Ω – ЦПНЗ, содержащее все множество элементов цифрового научного пространства, размещенных в некотором (возможно, распределенном) хранилище. Оно включает, в свою очередь, *n* подпространств, каждое из которых относится к определенной области науки. Каждое подпространство $\boldsymbol{\Omega}_{\mathrm{i}}$ состоит из двух множеств. Первое из них (обозначим его A_{i}) состоит из пронумерованных неко-

торым образом цифровых образов объектов реального мира (оцифрованные публикации, архивные документы, фотографии и пр.) и объектов, созданных исключительно в цифровой среде (электронные публикации, 3D-модели, мультимедийные материалы и т. п.). Нумерация должна однозначно идентифицировать объект и обеспечивать возможность его извлечения из хранилища. Второе множество (обозначим его B_{i}) включает метаданные, содержащие многоаспектные характеристики объектов первого множества, обеспечивающие их выборку по запросам к ЦПНЗ и представление пользователям.

Множество A_{i} состоит из элементов a_{ij}, где *j=1…N* (*N* – общее количество объектов, отраженных в Ω_{i}). В качестве этих элементов выступают объекты следующих видов:

- текстовые файлы (распознанные оцифрованные печатные или рукописные документы) или документы, изначально сформированные в электронном виде;
- статические изображения (нераспознанные оцифрованные документы, оцифрованные или изначально сформированные в цифровом виде фотографии);
- цифровые или оцифрованные аудиозаписи;
- цифровые или оцифрованные видео/киноматериалы;
- 3D-модели различных предметов;
- мультимедийные инсталляции (цифровые модели природных процессов и технических устройств, учебные материалы, виртуальные экскурсии и т. п.).

Если элементы множества A_{i} представлены простой совокупностью пар «объект – его номер», то множество B_{i}, в общем случае, представляет собой достаточно сложную фасетно-иерархическую структуру. Каждый его элемент представлен не только конкретным значением и ссылкой на элемент множества a_{ij} (что имеет место в традиционных библиографических информационно-поисковых системах), но может включать указание на связи с другими элементами. Таким образом, под элементами множества B_{i} будем понимать структуру, включающую смысловое значение характеристики объекта, указания на один или несколько элементов множества A_{i}, к которому относится данная характеристика, и

указание на связи с другими структурами, являющимися также элементами множества B_{i}.

В качестве составляющих элементов множества B_{i} могут выступать индексы классификационных систем (таких, как ГРНТИ, УДК и пр.), отражающих тематику документов, индивидуальные характеристики персоны (фамилия и имя, дата рождения и т. п.), наименования событий, их текстовые описания, временные и географические характеристики объектов и др.

Для обеспечения точности поиска объектов в ЦПНЗ множество B_{i} должно включать ряд непересекающихся подмножеств, характеризующих различные аспекты информации об элементах множества A_{i}. Очевидно, что таких разбиений может быть бесконечно много, но ограничимся рассмотрением «интуитивно-минимального» набора данных (но охватывающего широкий спектр характеристик объектов), включающего классы типа «что (кто), где, когда» (подмножество $B_{\mathrm{i}3}$,), дополненного классом «тематика» (подмножество $B_{\mathrm{i}4}$), формальными характеристиками, специфичными для ЦПНЗ, выделенными в подмножества $B_{\mathrm{i}1}$ (виды объектов, перечисленные выше) и $B_{\mathrm{i}2}$ (условия предоставления пользователям тех или иных объектов множества A_{i}).

Подмножество $B_{\mathrm{i}1}$ множества B_{i} состоит из 6-ти элементов, совпадающих с перечисленными выше:

b_{i11} – текстовый вид с возможностью поиска фрагмента текста;

$b_{\mathrm{i}12}$ – статическое изображение;

$b_{\mathrm{i}13}$ – 3D-объект;

$b_{\mathrm{i}14}$ – аудиодокумент;

$b_{\mathrm{i}15}$ – видеодукумент;

$b_{\mathrm{i}16}$ – – мультимедийный объект.

Подмножество $B_{\mathrm{i}2}$ множества B_{i} состоит из элементов, определяющих условия предоставления цифрового объекта пользователю. Введение данного подмножества обусловлено различными законодательными требованиями к публичному представлению объекта. Элементы множества B_2:

$b_{\mathrm{i}21}$ – объект находится в свободном доступе;

$b_{\mathrm{i}22}$ – объект находится в доступе, бесплатном для определенной группы

пользователей (например, оплаченная подписка на полнотекстовые научные издания для сотрудников некоторого учреждения) и недоступном для остальных пользователей;

$b_{\mathrm{i}23}$ – объект находится в ограниченном доступе, бесплатном для определенной группы и коммерческом для остальных пользователей (например, цифровая модель музейного экспоната может быть доступна для бесплатного просмотра посетителям музея, а удаленный просмотр предусматривает определенную плату;

$b_{\mathrm{i}24}$ – объект находится в коммерческом доступе, т. е. пользователю необходимо оплатить доступ к данному ресурсу.

Подмножество B_{i3} множества B_{i} включает: обозначение **типа** объекта (персона, организация, публикация, архивный документ, минерал, теорема и т. п., в зависимости от направления науки); основные характеристики объекта, необходимые для его идентификации при поиске («**что**» или «**кто**», «**где**», «**когда**»); **условия визуализации**. В качестве обязательного элемента множества B_{i3}, относящегося к классу «**что (кто)**», выступает название конкретного объекта (имя персоны), которое может быть дополнено элементами, конкретизирующими вид объекта внутри данного типа (например, для типа «публикация» это могут быть варианты: «научная монография», «научная статья», «поэтический сборник», учебник и т. д.), а также неструктурированными пояснениями, содержащими ту или иную информацию об объекте. Это может быть биография ученого, аннотация публикации, описание музейного предмета и т. п.). Например, коллекция фотографий Москвы 1930-х годов может быть дополнена развернутой статьей об архитектуре города того времени, представленной в виде гипертекста.

В качестве элемента класса «**где**» множества B_{i3} могут выступать различные реалии, связанные как непосредственно с географической принадлежностью объекта (например, для персоны – место рождения, для музейного объекта – место его первоначального обнаружения, для события – страна или город, где оно произошло, и т. п.), так и с организацией, описанной, в свою очередь, своими метаданными (например, места работы персоны, место хранения музейного предмета, издательство для печатного документа и т. п.).

В качестве элемента класса «**когда**» множества B_3 может выступать, напри-

мер, год публикации печатного издания, год рождения персоны, дата запуска космического корабля и т. п.

Класс **«условия визуализации»** содержит информацию о группах пользователей, которым может предоставляться данный объект без каких-либо условий, и условия предоставления объекта другим группам пользователей

Отметим, что подмножества B_{i1}, B_{i2} и B_{i3} множества B_{i} не пересекаются между собой.

Подмножество B_{i4} содержит элементы класса «тематика», оно может иметь достаточно сложную структуру, содержащую индексы и наименования элементов различных классификационных систем – строго иерархическую типа ГРНТИ [7], фасетную – типа УДК [8] и т. п.), ключевые термины, в том числе, оформленные в виде тезаурусов.

2. МОДЕЛЬ ОБРАБОТКИ ЗАПРОСОВ В ЦПНЗ

Для упрощения дальнейшего изложения будем предполагать, что мы рассматриваем подпространство ЦПНЗ, относящееся к одной из научных областей (если не будет сказано иного), и опустим нижний индекс при рассмотрении множеств A_{i} и. B_{i} Соответственно, при обозначении подмножеств этих множеств перейдем от двойных индексов к одинарным. Обозначим множество запросов пользователей, в соответствии с которыми из ЦПНЗ отбираются интересующие их документы, через F. Его составляющие (f_s) могут содержать элементы естественного языка, рубрики тех или иных классификационных систем, химические формулы, математические выражения и т. п., связанные булевыми операторами. Это множество, в отличие от конечных множеств A и B, содержит бесконечное число элементов ($F = \bigcup_s^{\infty} f_s$). Его элементами являются как разовые запросы отдельных пользователей, так и постоянные запросы, формируемые в рамках систем избирательного распространения информации [10, 11], а также запросы, в соответствии с которыми в ЦПНЗ формируются те или иные коллекции документов.

Если некоторый запрос f_s удовлетворяет условию $f_s = \bigcup_n^N f_{sn}$, где $f_{sn} \in B$ (т. е. f_{sn} представляет собой логическое выражение, включающее элементы одного из подмножеств B_{i}), то задача выбора и визуализации объектов из ЦПНЗ сводится к сравнению составляющих запроса f_{sn} с элементами подмножеств B_{i} (i=1,3,4) (назовем это линейным поиском). Результатом сравнения являются

адреса соответствующих элементов множества A, по которым эти элементы извлекаются из хранилища и предоставляются пользователю в соответствии с условиями, отраженными в подмножестве B_2.

Однако на практике запросы пользователя зачастую либо пересекаются с множеством *B* лишь частично, либо вообще не пересекаются. Это приводит к тому, что результат линейного поиска содержит лишь часть объектов ЦПНЗ, необходимых пользователю, либо не содержит их вообще, хотя во множестве A они имеются. Например, пользователю необходима подборка произведений поэтов «Серебряного века», отраженных в электронной библиотеке (ЭБ) публикаций XX века. При формировании элементов ЭБ, с большой долей вероятности, «Серебряный век» не указывался в качестве временной характеристики, его также нет ни в одной из классификационных систем, которые могли использоваться при формировании ЭБ. Соответственно, если обработать запрос в терминах «поэты Серебряного века», его результатом будет пустое подмножество элементов множества A. В то же время, очевидно, что среди элементов множества A есть объекты, соответствующие требованиям, предъявляемым к данной коллекции. Для их обнаружения необходимо построить отображение пользовательского запроса на множество B и далее реализовать линейный поиск по запросу, включающему соответствующие элементы подмножеств B_1 и B_3.

3. ФОРМИРОВАНИЕ ПОЛЬЗОВАТЕЛЬСКИХ КОЛЛЕКЦИЙ

Под пользовательской коллекцией будем понимать коллекцию элементов ЦПНЗ или электронной библиотеки, соответствующих запросу, сформулированному на естественном языке.

Следуя принципам формализации общего подхода к формированию пользовательских коллекций, построим иерархию их представления в ЦПНЗ.

На первом уровне этой иерархии располагаются элементы множества A, отобранные в соответствии с запросом, отображаемым на подмножество B_3 (предметные коллекции); на втором – элементы множества A, соответствующие запросу, отображаемому на подмножества B_1 и B_3 (предметно-видовые коллекции); третий уровень (тематико-видовые коллекции) формируется по запросам, отражаемым на подмножества B_1 и B_4; на четвертом уровне располагаются тематические коллекции, соответствующие запросам, отражаемым на подмножество

B_4. Наконец, на самом верхнем уровне иерархии располагаются междисциплинарные коллекции. Запросы на формирование таких коллекций не могут быть отражены с необходимой полнотой на одном множестве B, соответствующем той или иной области науки. Для получения полной коллекции объектов, соответствующих такому запросу, необходимо рассматривать несколько подпространств ЦПНЗ и строить отражение запроса на соответствующие подмножества нескольких множеств B_i.

В качестве примера коллекции первого уровня можно привести подборку всех материалов, касающихся конкретного ученого. Например, на запрос «М.В. Ломоносов» будут выданы произведения М.В. Ломоносова; связанные с ним публикации; МГУ им. М.В. Ломоносова; хребет Ломоносова и т. д.

Коллекция опубликованных трудов М.В. Ломоносова относится ко второму уровню иерархии. Отражение такого запроса на подмножество B_1 предписывает отбирать объекты вида b_{11} и b_{12}. Отражение запроса на подмножество B_3 позволяет осуществить линейный поиск по условию: «выбрать тип объекта «персона» с фамилией Ломоносов, инициалами М. В. (класс характеристик **«Кто»**) и тип объекта «публикация», в авторах которых (класс характеристик **«Кто»**) указаны выбранные персоны. В результате будет получен ряд полных текстов публикаций, автором которых является М.В. Ломоносов. Однако, в силу того, что в ЭБ или в ЦПНЗ могут быть отражены несколько персон с «именем» М.В. Ломоносов, сформированная коллекция будет содержать все их публикации. Если при запросе на формирование коллекции подразумевался конкретный Михаил Васильевич Ломоносов, родившийся в 1711 году, полученный результат будет некорректным. Чтобы получить требуемую коллекцию, отражение запроса на множество B_3 должно содержать год рождение персоны (класс характеристик **когда**).

В качестве примера пользовательской коллекции третьего уровня можно привести объекты, извлеченные из ЦПНЗ по запросу «3D-модели антропологических объектов». Для получения такой коллекции этот запрос необходимо отразить на подмножества B_1 и B_4. Первое предписывает отбирать объекты вида b_{13}, а для получения второго необходимо анализировать конкретные классификационные системы, используемые в ЭБ (ЦПНЗ).

К пользовательской коллекции верхнего уровня можно отнести такие коллекции, как «материалы, связанные с освоением космического пространства», в

которые должны быть включены объекты, относящиеся к физике, механике, технике, химии, астрономии, возможно, к философии, политологии и т. п.

Создание такой иерархии позволяет оптимизировать процесс формирования и сопровождения информационных фондов электронных библиотек, а также позволяет пользователю выбрать из всего множества взаимосвязанных ресурсов электронной библиотеки те информационные объекты, которые объединены одним или несколькими признаками.

В соответствии с моделью алгоритм формирования пользовательской коллекции включает следующие этапы:

1. Анализ соответствия терминов, включенных в запрос f_s, элементам множества B, определение уровня иерархии, которому соответствует пользовательская коллекция.
2. Разбиение терминов запроса на два подмножества: подмножество, содержащее в явном виде элементы множества B (например, вид объекта, тип объекта и т. п.) $- f_{s1}$, и подмножество, не содержащее в явном виде элементы $B\ - f_{s2}$.
3. Реализация алгоритма отображения элементов подмножества f_{s2} на множество B; формирование запроса $f_s^{\ b} \in B$.
4. Линейный поиск элементов множества *A,* отвечающих запросу $f_s^{\ b}$.
5. Формирование пользовательской коллекции в соответствии с условиями визуализации.

В качестве примера реализации описанного алгоритма рассмотрим формирование коллекции материалов, относящихся к поэтам «Серебряного века», в среде некоторой условной электронной библиотеки, организованной по принципам ЭБ «Научное Наследие России» (ЭБ ННР) [12].

Пусть имеется некоторая ЭБ *X*, отражающая культурное наследие России в части, относящейся к литературным произведениям. В качестве объектов ЭБ X выступают персоны (авторы) и публикации. Метаданные персон включают следующие элементы.

Фасет **КТО:**

- фамилия;
- имя;
- отчество;

- варианты имени (псевдонимы).

Фасет **ГДЕ:**

- место рождения;
- место смерти.

Фасет **КОГДА:**

- дата рождения;
- дата смерти;

Фасет **ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ:**

биография;

библиография.

Метаданные публикаций включают следующие элементы:

Фасет **ЧТО:**

- название публикации

- вид представления публикации (ссылка на элемент множества B_1);
- тип публикации (проза, поэзия, литературоведческая работа);
- вид публикации (монография, сборник, том многотомника, выпуск сериального издания, статья из сборника или из сериального издания, прочие виды);

Фасет **КТО:**

- ссылки на персон (авторов);
- ссылки на персон (редакторов, составителей, художников).

Фасет **ГДЕ:**

- место издания (страна, город, издательство);
- ссылка на публикацию для статей из журналов, сборников и т. п.;
- информация о конкретном выпуске издания (журнала, сборника), в котором опубликован данный материал;

Фасет **КОГДА:**

- год издания публикации.

- Фасет **ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ:**

полное библиографическое описание материала;

аннотация;

примечания.

Фасет **ТЕМАТИКА:**

индексы УДК;
индексы ББК;
ключевые термины.

Публикации, представленные в виде аудиозаписей, в фасете **КОГДА** должны содержать информацию о дате создания звукозаписи, а в фасете **ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ** (или в фасете **КТО**, если участники создания звукозаписи входят в круг персон, отраженных в ЭБ Х) – информацию о ее создателях.

Задача состоит в том, чтобы сформировать пользовательскую коллекцию материалов, входящих в ЭБ Х, относящихся к поэтам «Серебряного века». Коллекция должна включать сведения об авторах и обо всех документах, связанных с ними (в том числе, выходные данные и полные тексты их произведений).

Для решения этой задачи определим сначала, к какому временному интервалу относится понятие «Серебряный век».

Согласно «Литературной энциклопедии» [13], «Серебряный век» заключен в интервале между 1890-м и 1921-м годами.

Формирование искомой пользовательской коллекции включает следующие этапы:

- На первом этапе определяются годы публикаций (отражение на множество B), тем самым «переводится» параметр запроса «Серебряный век» на «язык» метаданных.
- На втором этапе необходимо выбрать из множества объектов «публикация» те, в фасете **ЧТО** метаданных которых имеется элемент «тип публикации», содержащий значение «поэзия», а в фасете **КОГДА** – один из годов, входящих в заданный интервал. Необходимо отметить, что для упрощения задания интервалов годов поисковый интерфейс ЭБ (ЦПНЗ) должен предусматривать в фасете **КОГДА** возможность обработки логических выражений, содержащих условия «больше», «меньше», «равно», «не равно». Соответственно, программная оболочка ЭБ должна уметь обрабатывать такие условия. Публикации, найденные на этом этапе, включаются в пользовательскую коллекцию.
- На третьем этапе выбираются персоны, ссылка на которых имеется в фасете **КТО** выбранных на предыдущем этапе публикаций. Данные об

этих персонах включаются в пользовательскую коллекцию.

- На четвертом этапе осуществляется визуализация сформированной коллекции в соответствии с условиями по каждому объекту. Программные средства ЭБ (ЦПНЗ) должны обеспечивать гибкие возможности визуализации элементов пользовательской коллекции сортировку по различным элементам метаданных различных объектов (это может быть не только список объектов, упорядоченный по алфавиту (или числовому значению) заданного элемента метаданных, но и список, в котором чередуются объекты разного вида). В частности, для рассматриваемого примера это может быть информация о поэте, за которой следует список его произведений со ссылками на полные тексты.

4. ЗАКЛЮЧЕНИЕ

Предлагаемые подходы к формализации процессов представления элементов ЦПНЗ и построения пользовательских коллекций позволяют упростить алгоритмизацию построения отдельных элементов ЦПНЗ, разработки его программной оболочки и пользовательского интерфейса. Работы в этом направлении проводятся в рамках государственного задания.

СПИСОК ЛИТЕРАТУРЫ

1. *Антопольский А.Б., Каленов Н.Е., Серебряков В.А., Сотников Н.А.* Точка зрения о едином цифровом пространстве научных знаний // Вестник Российской академии наук, 2019 (в печати).
2. *Gauch S., Chaffee J., Pretschner A*. Ontology-based personalized search and browsing. // Web Intell Agent Syst. 2003. V. 1. No 3, 4. P. 219–234.
3. *Sun Y., Yu Y., Han J.* Ranking-based clustering of heterogeneous information networks with star network schema // KDD '09 Proceedings of the 15th ACM SIGKDD international Conference on Knowledge discovery and data mining. 2009. P 797–806.
4. *Wong W., Liu W., Bennamoun M.* Ontology learning from text: a look back and into the future // ACM Computing Surveys (CSUR). 2012. V. 44. Issue 4. Article No 20.

5. *Chi Wang, Jialu Liu, Nihit Desai, Marina Danilevsky, Jiawei Han*. Constructing topical hierarchies in heterogeneous information networks // Knowledge and Information Systems. 2015. V. 44. Issue 3. P. 529–558.
6. *Каленов Н.Е., Соболевская И.Н., Сотников А.Н.* Иерархические уровни представления информационных объектов в среде электронных библиотек // Информация и инновации. 2018. Т. 13. № 2. С. 25–31.
7. *Антопольский А.Б., Белоозеров В.Н., Маркарова Т.С., Дмитриева Е.Ю.* Установление соответствий рубрик ГРНТИ рубрикам других систем классификации научной и технической информации // Научно-техническая информация. Серия 1: организация и методика информационной работы. 2015. № 3. С. 3–18.
8. *Астахова Т.С.* Проблемы отражения современного научного знания в классификационных системах: новое в УДК // Сборник трудов конференции «Перспективные направления научных исследований и критические технологии в классификационных системах» / ВИНИТИ РАН, Москва, 25–27 октября 2017 г. С. 32–35.
9. *Александров П.С.* Введение в теорию множеств и общую топологию. М.: «Наука», 1977. 368 с.
10. *Ивановский А.А.* Объектная модель системы избирательного распространения информации // Научные и технические библиотеки. 2019. № 4. С. 61–75. DOI 10/33186/1027-3689-2019-4-61-75
11. *Захарова С.С.* Избирательное распространение информации и информационно-коммуникационные технологии: обзор исследований // Библиотековедение. 2017. № 6. С. 651–658. DOI: 10.25281/0869-608X-2017-66-6-651-658
12. *Каленов Н.Е., Савин Г.И., Серебряков В.А., Сотников А.Н.* Принципы построения и формирования электронной библиотеки «Научное наследие России» // Программные продукты, системы и алгоритмы. Электронный журнал. 2012. Т. 4. № 100. С. 30– 40. Url: http://www.swsys-web.ru
13. *Литературная энциклопедия* [Электронный ресурс]. (https://dic.academic.ru/dic.nsf/enc_literature/5383/%D0%A1%D0%B5%D1%80%D0%B5%D0%B1%D1%80%D1%8F%D0%BD%D1%8B%D0%B9) (07.11.2019).

FORMALIZATION OF PROCESSES FOR FORMING USER COLLECTIONS IN THE DIGITAL SPACE OF SCIENTIFIC KNOWLEDGE

N. E. Kalenov, I. N. Sobolevskaya, A. N. Sotnikov

Joint Supercomputer Center of the Russian Academy of Sciences - Branch of Federal State Institution "Scientific Research Institute for System Analysis of the Russian Academy of Sciences"

nkalenov@jscc.ru, ins@jscc.ru, ansotnikov@jscc.ru

Abstract

The task of forming a digital space of scientific knowledge (DSSK) is analyzed in the paper. The difference of this concept from the general concept of the information space is considered. DSSK is presented as a set containing objects verified by the world scientific community. The form of a structured representation of the digital knowledge space is a semantic network, the basic organization principle of which is based on the classification system of objects and the subsequent construction of their hierarchy, in particular, according to the principle of inheritance. The classification of the objects that make up the content of the DSSK is introduced. A model of the central data collection system is proposed as a collection of disjoint sets containing digital images of real objects and their characteristics, which ensure the selection and visualization of objects in accordance with multi-aspect user requests. The concept of a user collection is defined, and a hierarchical classification of types of user collections is proposed. The use of the concepts of set theory in the construction of DSSK allows you to break down information into levels of detail and formalize the algorithms for processing user queries, which is illustrated by specific examples.

Keywords*:* recursive link, knowledge cyberdomain, digital library, detail levels, data entries hierarchy.

REFERENCES

1. *Antopol'skij A.B., Kalenov N.E., Serebryakov V.A., Sotnikov N.A*. Tochka zreniya o edinom cifrovom prostranstve nauchnyh znanij // Vestnik Rossijskoj akademii nauk, 2019 (v pechati).

2. *Gauch S., Chaffee J., Pretschner A*. Ontology-based personalized search and browsing. // Web Intell Agent Syst. 2003. V. 1. No 3, 4. P. 219–234.
3. *Sun Y., Yu Y., Han J.* Ranking-based clustering of heterogeneous information networks with star network schema // KDD '09 Proceedings of the 15th ACM SIGKDD international Conference on Knowledge discovery and data mining. 2009. P 797–806.
4. *Wong W., Liu W., Bennamoun M.* Ontology learning from text: a look back and into the future // ACM Computing Surveys (CSUR). 2012. V. 44. Issue 4. Article No 20.
5. *Chi Wang, Jialu Liu, Nihit Desai, Marina Danilevsky, Jiawei Han*. Constructing topical hierarchies in heterogeneous information networks // Knowledge and Information Systems. 2015. V. 44. Issue 3. P. 529–558.
6. *Kalenov N.E., Sobolevskaya I.N., Sotnikov A.N.* Ierarhicheskie urovni predstavleniya informacionnyh ob"ektov v srede elektronnyh bibliotek // Informaciya i innovacii. 2018. T. 13. No 2. S. 25–31.
7. *Antopol'skij A.B., Beloozerov V.N., Markarova T.S., Dmitrieva E.YU*. Ustanovlenie sootvetstvij rubrik GRNTI rubrikam drugih sistem klassifikacii nauchnoj i tekhnicheskoj informacii // Nauchno-tekhnicheskaya informaciya. Seriya 1: organizaciya i metodika informacionnoj raboty. 2015. No 3.S. 3–18.
8. *Astahova T.S.* Problemy otrazheniya sovremennogo nauchnogo znaniya v klassifikacionnyh sistemah: novoe v UDK // Sbornik trudov konferencii «Perspektivnye napravleniya nauchnyh issledovanij i kriticheskie tekhnologii v klassifikacionnyh sistemah». VINITI RAN, Moskva, 25–27 oktyabrya 2017 g. S. 32–35.
9. *Aleksandrov P.S.* Vvedenie v teoriyu mnozhestv i obshchuyu topologiyu. M.: «Nauka», 1977. 368 s.
10. *Ivanovskij A.A.* Ob"ektnaya model' sistemy izbiratel'nogo rasprostraneniya informacii // Nauchnye i tekhnicheskie biblioteki, 2019. № 4. S. 61–75.
11. *Zaharova S.S.* Izbiratel'noe rasprostranenie informacii i informacionno-kommunikacionnye tekhnologii: obzor issledovanij // Bibliotekovedenie. 2017. No 6. S. 651–658.
12. *Kalenov N.E., Savin G.I., Serebryakov V.A., Sotnikov A.N.* Principy postroeniya i formirovaniya elektronnoj biblioteki "Nauchnoe nasledie Rossii" //

Programmnye produkty, sistemy i algoritmy. Elektronnyj zhurnal. 2012. T. 4. No 100. S. 30–40. Url: http://www.swsys-web.ru

13. Literary encyclopedia [digital resource]. Url: https://dic.academic.ru/dic.nsf/enc_literature/5383/%D0%A1%D0%B5%D1%80%D0%B5%D0%B1%D1%80%D1%8F%D0%BD%D1%8B%D0%B9 (07.11.2019)

СВЕДЕНИЯ ОБ АВТОРАХ

КАЛЕНОВ Николай Евгеньевич – главный научный сотрудник Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», д. т. н., профессор. Сфера научных интересов – математическое обеспечение, программные средства и системы для распределенных вычислений; формирование баз данных для электронных библиотек; методы, средства и системы обработки данных большого объема.

Nikolay Evgenvich KALENOV – Chief Researcher of Joint Super Computer Center of the Russian Academy of Sciences – Branch of Federal State Institution “Scientific Research Institute for System Analysis of the Russian Academy of Sciences”. Research interests include mathematical software, software and systems for distributed computing; e-library database building; methods, tools and systems of large data processing.

email: nkalenov@jscc.ru

СОБОЛЕВСКАЯ Ирина Николаевна – старший научный сотрудник Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», к. ф.-м. н. Сфера научных интересов – математическое обеспечение, программные средства и системы для распределенных вычислений; формирование баз данных для электронных библиотек; методы, средства и системы обработки данных большого объема; 3D-моделирование.

Irina Nikolaevna SOBOLEVSKAYA – senior scientist researcher of Joint SuperComputer Center of the Russian Academy of Sciences – Branch of Federal State Institution "Scientific Research Institute for System Analysis of the Russian Academy of Sciences". Research interests include mathematical software, software and systems for distributed computing; e-library database building; methods, tools and systems of large data processing; 3D modeling.

email: ins@jscc.ru

СОТНИКОВ Александр Николаевич – зам. директора по научной работе Межведомственного Суперкомпьютерного Центра РАН – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», д. ф.-м. н., профессор. Сфера научных интересов – математическое обеспечение, программные средства и системы для распределенных вычислений; формирование баз данных для электронных библиотек; методы, средства и системы обработки данных большого объема; нейронные и семантические сети.

Aleksandr Nikolaevich SOTNIKOV – Deputy director for science of Joint SuperComputer Center of the Russian Academy of Sciences – Branch of Federal State Institution "Scientific Research Institute for System Analysis of the Russian Academy of Sciences". Research interests include mathematical software, software and systems for distributed computing; e-library database building; methods, tools and systems of large data processing; semantic and nerve nets.

email: ansotnikov@jscc.ru

Материал поступил в редакцию 16 ноября 2019 года