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

Математическое моделирование процессов формирования междисциплинарных коллекций в среде электронных библиотек

Рекомендуемая форма библиографической ссылки

Каленов Н.Е., Соболевская И.Н., Сотников А.Н. Математическое моделирование процессов формирования междисциплинарных коллекций в среде электронных библиотек // Научный сервис в сети Интернет: труды XXI Всероссийской научной конференции (23-28 сентября 2019 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2019. — С. 347-356. — URL: http://keldysh.ru/abrau/2019/theses/04.pdf doi:10.20948/abrau-2019-04

Размещена также презентация к докладу

Математическое моделирование процессов формирования междисциплинарных коллекций в среде электронных библиотек

Н.Е. Каленов[1], И.Н. Соболевская[1], А.Н. Сотников[1]

[1] Межведомственный суперкомпьютерный центр Российской академии наук – филиал Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», Москва, 119334 Москва, Ленинский проспект, 32а

Аннотация. В работе исследуется задача формирования цифрового пространства научных знаний (ЦПНЗ). Рассматривается отличие этого понятия от общего понятия информационного пространства. 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. Определено понятие иерархической связи между объектами. Использование понятий теории множеств при построении ЦПНЗ позволяет разбивать информацию по уровням детализации. Введено понятие уровней иерархии объектов ЦПНЗ. Даны определения объектов различных уровней иерархии и сформулированы принципы работы с объектами каждого из уровней. Показано как с помощью иерархической структуры представления информационных объектов в среде электронной библиотеки может быть сформирована пользовательская коллекция в ЦПНЗ Описаны методы обработки поисковых запросов в ЦПНЗ. Построение иерархий, представляющих «объектные» разделы с высокой степенью детализации, позволяет повысить эффективность поиска информации в пространстве знаний и анализа информации.

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

Math modeling of the interdisciplinary collections generation process in the digital libraries environment

N.E. Kalenov1, I.N. Sobolevskaya1, A.N. Sotnikov1

1 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, 119334 Moscow, Leninsky Prospect, 32a

Abstract. The article deals with the problem of building a modern information society as the task of forming a knowledge cyberdomain. The implementation of this task can be carried out on the technological platform of the digital library, which ensures the formation and provision of information resources in various knowledge parts to general public users. The article presents the concept of "knowledge cyberdomain". The difference between the knowledge cyberdomain and the information space is presented in the paper. The concepts of data entries and types of data entries that make up the content of the knowledge cyberdomain are examines. Knowledge cyberdomain is presented in the form of a set containing data entries that have been tested by the world scientific community. Data entries properties of this set are transferred to all objects of subsets of this set, which allows to avoid a significant part of duplication of information in the knowledge cyberdomain. The objects linking to each other is defined as a recursive transition network. The presentation of the concepts of a set and a subset of a given set for building a digital knowledge space allows you to divide information into levels of detail. Introduced the concept of hierarchical relationships between data entries. The construction of hierarchies that represent "data entry" sections with a high degree of detail allows to increase the efficiency of information retrieval in the knowledge space and analysis of information. The concept of hierarchy level of data entries is presented in the article. Definitions of various levels data entries are given and principles of work with data entries of each level are formulated. The hierarchy of representation of electronic data entries in the knowledge cyberdomain is proposed.

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

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 и каталоги.

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 представляют 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»), с текстом этой публикации, с «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., обеспечивающим 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 данной работе мы рассматриваем ЦПНЗ в аспекте теории множеств, что позволяет подойти к вопросам построения пространства и работы с ним с новой точки зрения.

Пусть Ω – ЦПНЗ, содержащее все множество элементов цифрового научного пространства, размещенных в некотором (возможно, распределенном, хранилище). Оно включает, в свою очередь, два множества. Первое из них (обозначим его A) состоит из пронумерованных некоторым 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) включает метаданные, содержащие многоаспектные характеристики объектов первого множества, обеспечивающие их выборку по запросам к ЦПНЗ и представление пользователям.

Множество A состоит из элементов a_i, где $i=1…N$ (N- общее количество объектов, отраженных в Ω). 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 представлены простой совокупностью пар «объект – его номер», то множество B, в общем случае, представляет собой достаточно сложную фасетно-иерархическую структуру. Каждый его элемент представлен не только конкретным значением и ссылкой на элемент множества A (что имеет место в традиционных библиографических информационно-поисковых системах), но может включать указание на связи с другими элементами. Таким образом, под элементами множества B будем понимать структуру, включающую смысловое значение характеристики объекта, указания на один или несколько элементов множества A, к которому относится данная характеристика, и указание на связи с другими структурами, являющимися также элементами множества B.

В качестве составляющих элементов множества B 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 т.п.

Для обеспечения точности поиска объектов в ЦПНЗ множество B должно включать ряд непересекающихся подмножеств, характеризующих различные аспекты информации об элементах множества А. Очевидно, что таких разбиений может быть бесчисленное множество, но ограничимся рассмотрением «интуитивно-минимального», но охватывающего широкий спектр характеристик объектов, набора данных, включающего классы типа «что (кто), где, когда», дополненного классом «тематика» и формальными характеристиками, специфичными для ЦПНЗ, выделенными в подмножества B_1 (виды объектов, перечисленные выше) и B_2 (условия предоставления пользователям тех или иных объектов множества А).

Подмножество B_1 множества B ($B_1: B_1 \subset B$), состоит из 6-ти элементов, являющихся характеристиками элементов множества метаданных, которые назовем видами представления цифрового объекта. А именно:

b_{11} – текстовый вид с возможностью поиска фрагмента текста;

b_{12} – статическое изображение;

b_{13} – $3D$ – объект;

b_{14} – аудиодокумент;

b_{15} – видеодокумент;

b_{16} – мультимедийный объект.

Подмножество B_2 множества B ($B_2: B_2 \subset B$) состоит из элементов, определяющих условия предоставления цифрового объекта пользователю. Введение данного подмножества обусловлено различными законодательными требованиями к публичному представлению объекта. Элементы множества B_2 будем называть условиями предоставления объекта. А именно:

b_{21} – объект находится в свободном доступе;

b_{22} – объект находится в ограниченном доступе, бесплатном для определенной группы пользователей (например, оплаченная подписка на полнотекстовые научные издания для сотрудников некоторого учреждения) и, недоступном для остальных пользователей;

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

Подмножество B_3 множества B ($B_3: B_3 \subset B$) содержит основные характеристики объекта необходимые для его идентификации при поиске («**что**» или «**кто**», «**где**», «**когда**»). В качестве элемента множества B_3 ,

относящегося к классу «**что**», выступать в качестве обязательного элемента название конкретного объекта, которое может быть дополнено элементами, уточняющими тип объекта внутри данного вида, как то: книга, статья, архивный документ и т.п., а также неструктурированными пояснениями, содержащими ту или иную информацию об объекте. Например, коллекция фотографий Москвы 30-х годов может быть дополнена развернутой статьей об архитектуре города того времени, представленной в виде гипертекста.

В качестве элемента класса «**где**» множества B_3 могут выступать различные реализации, 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.). Например, коллекция гербариев А.Н. Петунникова [http://www.gbmt.ru/ru/about/fund/fondovaya-kollektsiya-gerbariy/, http://e-heritage.ru/ras/view/person/general.html?id=49901007], предусматривает указание на географическое место сбора объектов коллекции.

В качестве элемента класса «**когда**» множества B_3 может выступать, например, год публикации печатного издания, или год рождения персоны и т.п.

Отметим, что подмножества B_1, B_2 и B_3 множества B – не пересекаются между собой.

Подмножество B_4 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 слова и термины. В том числе оформленные в виде тезаурусов и т.д.

Если запросы пользователя к ЦПНЗ включают только элементы множества B, то задача выбора и представления документов сводится к формированию условий выборки, состоящей из терминов запроса, связанных булевыми операторами. Поиск данных в пространстве B осуществляется путем сравнения его элементов с элементами запроса (назовем это – линейным поиском); результатом поиска являются адреса соответствующих элементов множества A, по котором эти элементы извлекаются из хранилища и предоставляются пользователю в соответствии с условиями, отраженными в множестве B_2.

Однако на практике часто пользователю необходимо сформировать коллекцию из элементов множества A по признакам, не допускающим явную формулировку в терминах множества B. Например, требуется сформировать пользовательскую коллекцию из произведений поэтов «Серебряного века», отраженных в электронной библиотеке (ЭБ) публикаций 20-го века. При формировании элементов ЭБ «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 задать 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 множества B.

Отметим, что свойства объектов пространства Ω переносятся на все объекты подмножеств этого множества, что позволяет избежать значительной части дублирования информации [9].

Пусть множество F – множество «тематик объекта». Под тематикой объекта (f_s, где $s = 1 \dots \infty$) понимается заданный параметр, по которому будет осуществляться объединение объектов множества A в пользовательскую коллекцию , например, таким параметром может выступать некоторое научное направление или «сущность» объекта (минерал, поэты серебряного века, минералы, упоминающийся в поэзии серебряного века и т.п.), т.е. $F = \bigcup_s^\infty f_s$. Применив некоторое отображение f_s к элементам множеств A и B, можно получить, так называемую, коллекцию – т. е. совокупность объектов одного или нескольких видов, объединенных заданным параметром: $f_s(B(A), A)$.

Тогда, объединение множеств A и F по некоторому общему признаку, определенному в Ω_1, назовем множеством коллекций G.

Введенные выше понятия и определения, позволяют построить иерархию представления цифровых объектов в среде электронной библиотеки, что, в свою очередь, допускает формализацию общего подхода к формированию пользовательских коллекций в ЦПНЗ (под которыми мы будем понимать совокупность объектов одного или нескольких видов и (или) тематик).

На первом уровне этой иерархии располагаются элементы множества A, на втором уровне будут находиться элементы $\mathcal{A}_k$, которые назовем ***тематико-видовыми*** коллекциями, на третьем уровне располагаются элементы $f_s(\mathcal{A}_k(A), A)$, которые будем называть ***тематическими коллекциями*** и, наконец на верхнем – четвертом уровне находятся элементы множества G. Эти элементы назовем ***междисциплинарными коллекциями***.

Т.е., множество $\Omega = \bigcup(A, \mathcal{A}_k, F, G)$, при этом:

$$A \subset \mathcal{A}_k$$
$$\mathcal{A}_k \subset F$$
$$F \subset G$$

Отметим, что элементы множества метаданных B входят неявно в каждое из подмножеств множества Ω.

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.

С помощью иерархической структуры представления информационных объектов в среде ЭБ может быть сформирована пользовательская коллекция в ЦПНЗ.

Таким образом, задача формирования коллекций по определенному признаку, сводится к следующим этапам:

1. Анализ соответствия элементов этого признака коллекции элементам множества B.
2. Разбиение признаков коллекции на два подмножества: подмножество, содержащее в явном виде элементы множества B (например, вид объекта, тип объекта и т.п.), и подмножество не содержащее в явном виде элементы B.
3. Реализация алгоритма отображения характеристики коллекции на множество B.

В качестве примера реализации алгоритма отображения коллекции на множество метаданных, рассмотрим формирование коллекции материалов, относящихся к поэтам «Серебряного века» в среде ЭБ «Научное Наследие России» (ЭБ ННР). Такая коллекция должна включать сведения об авторах и обо всех документах, связанных с ними (в том числе выходные данные и полные тексты их произведений).

На первом этапе определяются годы публикаций, тем самым, «переводя» «параметр» запроса «Серебряный век» на «язык» метаданных. После этого осуществляется выборка всех объектов, входящих в ЭБ ННР по заданному временному интервалу. Далее, из найденного массива данных выбираются персоны, которые соответствуют таким элементам метаданных, как «автор», связанных, в свою очередь, с изданиями, имеющими элемент метаданных «тип издания» со значением «поэзия». В результате мы получаем список фамилий поэтов «серебряного века».

Затем, из всех полученных на первом этапе объектов производится выборка всех материалов по параметру «автор».

Таким образом, будет получена коллекция всех объектов (в том числе, архивных документов, фотодокументов и т.п., если таковые имеются в ЭБ ННР), относящиеся к поэтам «Серебряного века».

Работа выполнена при поддержке РФФИ (проекты 17-07-00400 и 18-07-00893).

Литература

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, vol. 1, № 3, 4, 2003. - pp. 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. - pp 797-806.
4. Wong W., Liu W., Bennamoun M. Ontology learning from text: a look back and into the future // ACM Computing Surveys (CSUR) / vol. 44 Issue 4, Article № 20, August 2012.
5. Chi Wang, Jialu Liu, Nihit Desai, Marina Danilevsky, Jiawei Han. Constructing topical hierarchies in heterogeneous information networks // Knowledge and Information Systems / vol. 44, Issue 3, September 2015. - pp 529–558.
6. Каленов Н.Е., Соболевская И.Н., Сотников А.Н. Иерархические уровни представления информационных объектов в среде электронных библиотек // Информация и инновации / Т. 13, № 2, 2018. – 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 / № 3, 2015. – 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.

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, vol. 1, № 3, 4, 2003. - pp. 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. - pp 797-806.
4. Wong W., Liu W., Bennamoun M. Ontology learning from text: a look back and into the future // ACM Computing Surveys (CSUR) / vol. 44 Issue 4, Article № 20, August 2012.

5. Chi Wang, Jialu Liu, Nihit Desai, Marina Danilevsky, Jiawei Han. Constructing topical hierarchies in heterogeneous information networks // Knowledge and Information Systems / vol. 44, Issue 3, September 2015. - pp 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 / T. 13, № 2, 2018. – 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 / № 3, 2015. – 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.