Применение модели направленных связей между документами для построения информационных систем научного сообщества

© В.Б. Барахнин, Ю.В. Леонова

Институт вычислительных технологий СО РАН
bar@ict.nsc.ru, juli@ict.nsc.ru

Аннотация

В статье обсуждается информационная модель горизонтальных направленных связей между документами на основе бинарных отношений с дополнительными атрибутами. Рассматривается ее применение для построения научных информационных систем.

Введение

При создании информационно-справочных систем с доступом через Internet для научного сообщества представляется важным предоставить пользователю максимально удобную навигацию по системе, в полной мере использующую возможности гипертекста. К сожалению, эти возможности реализуются далеко не всегда. В частности, если рассматривать информационно-справочные системы математической направленности, то в наиболее известных отечественных разработках – порталах MathTree [1] и Math-Net.RU [2] – внутренние иерархические связи между документами отсутствуют. В некоторых зарубежных системах такие связи имеются, но эти системы, как правило, специализированы по жанру ресурса. Например, портал MacTutor History of Mathematics [3] содержит весьма подробную биографическую информацию с перекрестными ссылками, но библиографическая информация представлена в виде кратких списков трудов.

При проектировании информационных систем возникает проблема возможного рассогласования информации. Во-первых, включение в документы информации о разнородных сущностях может привести к появлению множественной информации об одном и том же объекте. Такая ситуация, возможна, например, когда человек работает в разных организациях, участвует в разных проектах, является автором множества публикаций. Это может вызвать серьезные проблемы в случае необходимости появления различных версий информации, возникающих вследствие ее модификации.

Кроме того, для представления сложных документов, когда один документ является частью другого (полностью или частично, в том числе и в виде гиперссылки), необходимо выработать подходы к установлению связей между документами. Такая ситуация возникает, если о сущностях, описываемых документами, может быть построено истинное высказывание (представляющее интерес с точки зрения содержания системы) типа: «*Сущность А есть* (или *была*) *нечто* (по отношению к) *сущности Б*» или «*Сущность А имеет* (или *имела*) *в некотором качестве сущность Б*». Например: «Евклид – автор «Начал»» или «С.Л.Соболев был директором Института математики СО РАН». Нетрудно видеть, что типы таких связей могут быть различными, и это обстоятельство нужно учитывать в процессе разработки модели отношений между документами.

Таким образом, становится актуальной разработка технологии идентификации, спецификации и визуализации горизонтальных отношений между сущностями, информация о которых содержится во множестве документов, а также между документами, которые являются составной частью сложных документов. Одним из основных элементов этой технологии является разработка информационной модели отношений и тематических связей между документами системы.

Отметим, что в библиотечных системах, построенных на основе протокола Z39.50 и его версий [4], выполняется полное дублирование служебной информации. Аналогичная ситуация возникает в информационных системах, построенных на основе LDAP-каталогов [5], в

Труды 7ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2005, Ярославль, Россия, 2005.

которых имеется мощная система перекрестных ссылок, но используемая иерархическая модель не допускает отношений «многие-ко-многим». Если такие отношения все же возникают, то появляется необходимость дублирования информации, что может привести к рассогласованию информации в системе.

Ввиду этого, в информационных системах, подобных разрабатываемой нами, целесообразно хранить информацию в единственном экземпляре, устанавливая в нужных случаях отношения «многие-ко-многим».

Разумеется, приемы решения подобной задачи обсуждались ранее в ряде работ, например, [6-8]. Однако основной подход к представлению данных в этих работах заключается в рассмотрении многоместных отношений с их последующей декомпозицией в процессе нормализации. Мы же строим информационную модель с использованием только бинарных отношений, приписывая им дополнительные атрибуты, не укладывающиеся в общую схему. Таким образом, декомпозиция проводится на более высоком уровне абстрагируемости от структуры данных, что делает нашу модель более универсальной.

Модель документа в системе

Информационная система представляет собой множество связанных различными отношениями документов, описывающих некие сущности (объекты, факты или понятия). Информация о той или иной сущности содержится в системе либо непосредственно в виде документа, который ее представляет, описывает или моделирует, либо в виде упоминаний об этой сущности, которые имеются в других документах, т.е. содержат опосредованную информацию об этой сущности. Непременным атрибутом информационной системы, отличающим ее от обычных веб-сайтов, является наличие каталога, в котором содержатся метаданные документов.

Согласно стандартам построения открытых систем (OSI) [9] структура и содержание документа должны описываться в соответствии с международными схемами данных. Для описания соответствующих схем данных используются метаданные, которые определяют структуру и смысловое содержание документа. В нашей системе документом называется информационный ресурс, снабженный метаописанием (метаданными) в соответствии с рекомендациями OSI.

Дадим два определения:

Определение 1: Документом d_i называется пара: $d_i = <S_i, V_i>$,

где S_i – структура документа в соответствии с выбранной схемой данных;

V_i – содержание документа (информационное наполнение).

Определение 2: Коллекция – множество документов с выделенной фиксированной структурой, содержание которых имеет одинаковую тематическую направленность.

С точки зрения унификации работы с документами будем представлять информационную систему в виде набора коллекций. Метаданные, описывающие структуру и содержание документов в коллекциях, подразделяются на описательные и структурные.

Структурные метаданные определяют структуру и свойства документов, в соответствии с которыми осуществляется их обработка (типы, связи, форматы представления, ограничения на управление доступом и т.п.).

Описательные метаданные описывают смысловое содержание документа (его название, краткое содержание и т.п.).

Отметим, что описательные метаданные, характеризующие документ, могут являться частью документа и в то же время могут содержать в соответствии с выбранной *схемой данных* сведения о документе (основные и дополнительные, такие, как, например, авторы, название, дата создания и т.д.).

Элемент схемы данных данной коллекции будем называть *структурным элементом*.

Структурный элемент (далее просто элемент) имеет идентификатор и обладает некоторыми свойствами.

Таким образом, элемент E – это совокупность $<ID, P>$,

где ID – идентификатор элемента,

P – свойства элемента.

Экземпляр элемента имеет значение (или содержание).

Свойства элемента определяют характер работы с элементом.

Элемент обладает типом, выбираемым из словаря. Тип определяет правила работы с элементом и, следовательно, является свойством элемента.

Примеры элементов: заголовок документа, аннотация документа, фамилия в визитной карточке, авторы документа. Значение элемента – его конкретная содержательная часть, а свойства элемента описывают его структуру. Для элемента визитной карточки «Фамилия» значение – Матвеев, идентификатор – 1, свойства – тип "word".

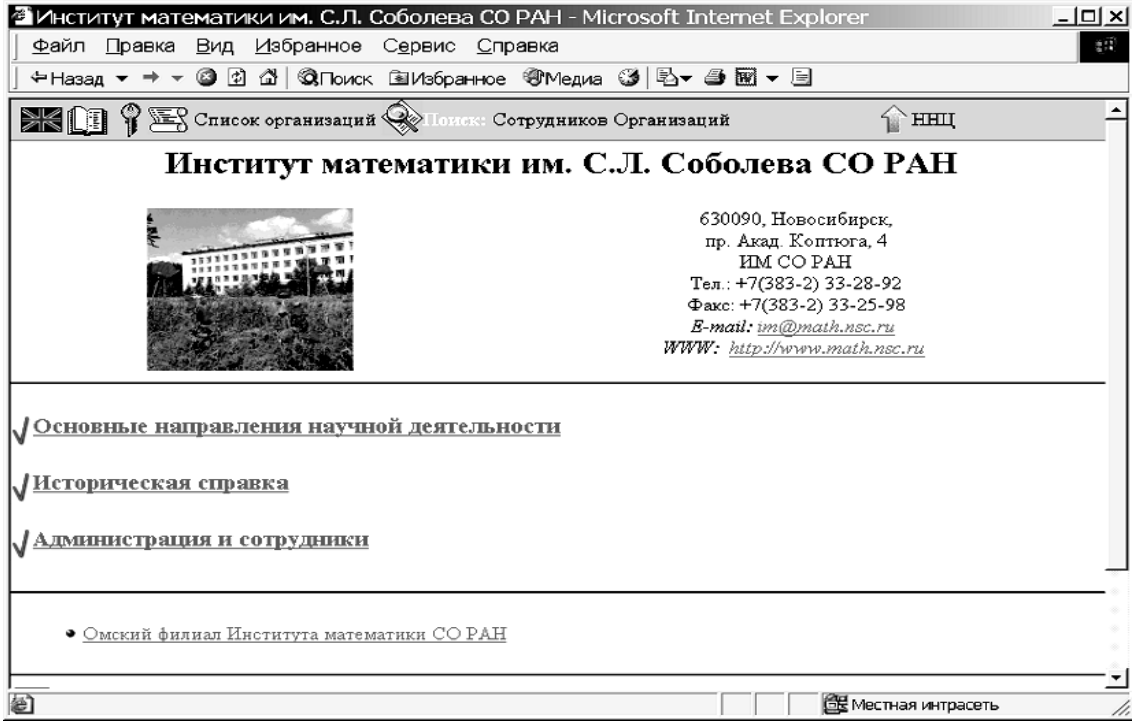

Рис.1. Пример навигационной панели

Отметим, что значением элемента может быть и внешний объект. Например, для элемента визитной карточки «Фотография» значением является внешний объект – графический файл.

Структура документа – это набор структурных элементов.

Содержание документа – объединение значений экземпляров элементов, составляющих документ.

Для решения сформулированных выше задач мы должны определить связи (отношения) между документами.

Модель отношений между документами в информационной системе

В основу нашей модели отношений между документами в информационной системе легла модель RDF [10], которая описывает ресурсы и отношения между ними. Описание ресурса в RDF – это совокупность утверждений о свойствах ресурса. Каждое утверждение представляет собой тройку: ресурс, именованное свойство и его значение. Отношения между ресурсами представляются именованными свойствами. Например, утверждение, что Коробейников С.Н. является автором книги “Нелинейное деформирование твердых тел” в терминах RDF в нотации N-Triples может быть выражено следующим образом:

```
<Коробейников С.Н.>
<AuthorOf>
<Нелинейное деформирование твердых тел>.
```

Основное отличие нашей модели от модели RDF состоит в том, что выстраиваемые нами отношения переносятся на уровень элементов, определяющих структуру документов. В рассматриваемой информационной системе само отношение определяется не ресурсом, а структурными метаданными коллекций документов системы.

В нашей системе связи между документами устанавливаются путем задания на множестве документов бинарных отношений, которые в соответствии с одной из форм нотации, используемых RDF, могут быть записаны в виде A(R,V): объект R имеет атрибут A со значением V. Например, тот факт, что Барахнин В.Б. занимает некоторую должность (post) в ИВТ СО РАН, записывается как `Post`(‘ИВТ СО РАН’, ‘Барахнин В.Б.’), где `Post` – то или иное значение из списка (тезауруса) должностей.

В информационно-справочных системах для научного сообщества мы выделяем два вида отношений:

- Отношение порядка между документами, выстраивающее иерархию подчинения в коллекции, например отношение подчиненности между документами в коллекции «Организации»:

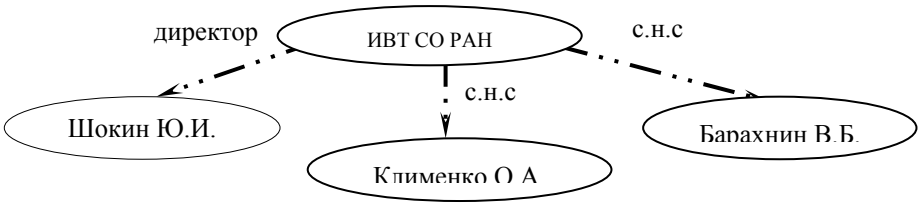

Рис.2. Связи между документами коллекции “Организации” и коллекции “Персоны”

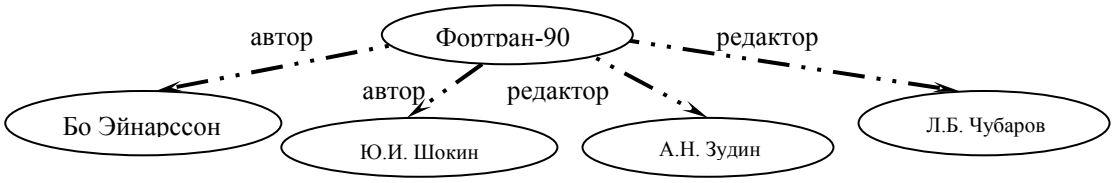

Рис.3. Связи между документами коллекции “Публикации” и коллекции “Персоны”

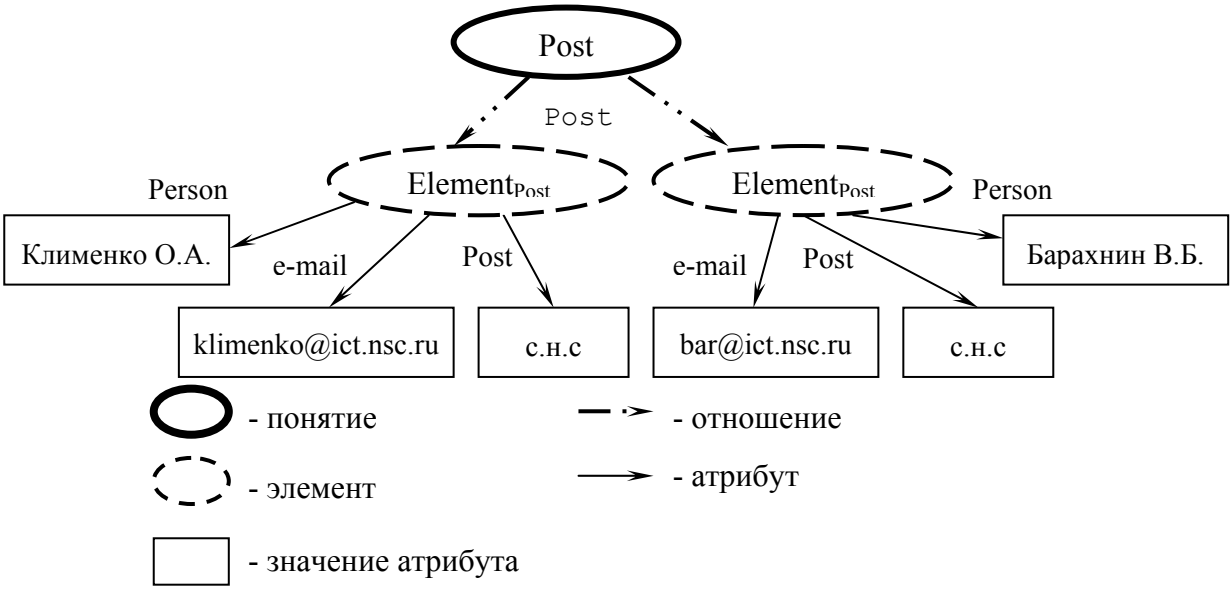

Рис. 4. RDF-представление дополнительных атрибутов отношений второго типа с учетом множественного типа

`Head` (‘Кафедра математического моделирования’, ‘ММФ НГУ’).
Отметим, что данный тип отношения предполагает установление только односторонней связи между документами.

– Отношение связи между документами, например отношение типа принадлежности между документами коллекции «Организации» и документами коллекции «Персоны»:
`Post` (‘ИВТ СО РАН’, ‘Барахнин В.Б.’).
Данный тип отношения допускает установление двусторонней связи между документами, в том смысле, что одновременно может существовать и A(R,V). Таким образом, любой объект также может играть и роль значения.

Различие отношений первого и второго типа заключается в том, что отношениям первого типа изначально приписано свойство – иерархия, а отношениям второго типа никаких свойств изначально не приписано. Свойства отношений второго типа определяются для каждого конкретного отношения.

Отношение первого типа, как правило, имеет не более одного атрибута, например тип подчинения (территориальное, научно-методическое и т.д.).

Отношение второго типа, как правило, имеют несколько дополнительных атрибутов. Например, отношение типа "`Post`" не просто описывает принадлежность персоны к организации, но и обладает следующими атрибутами: название

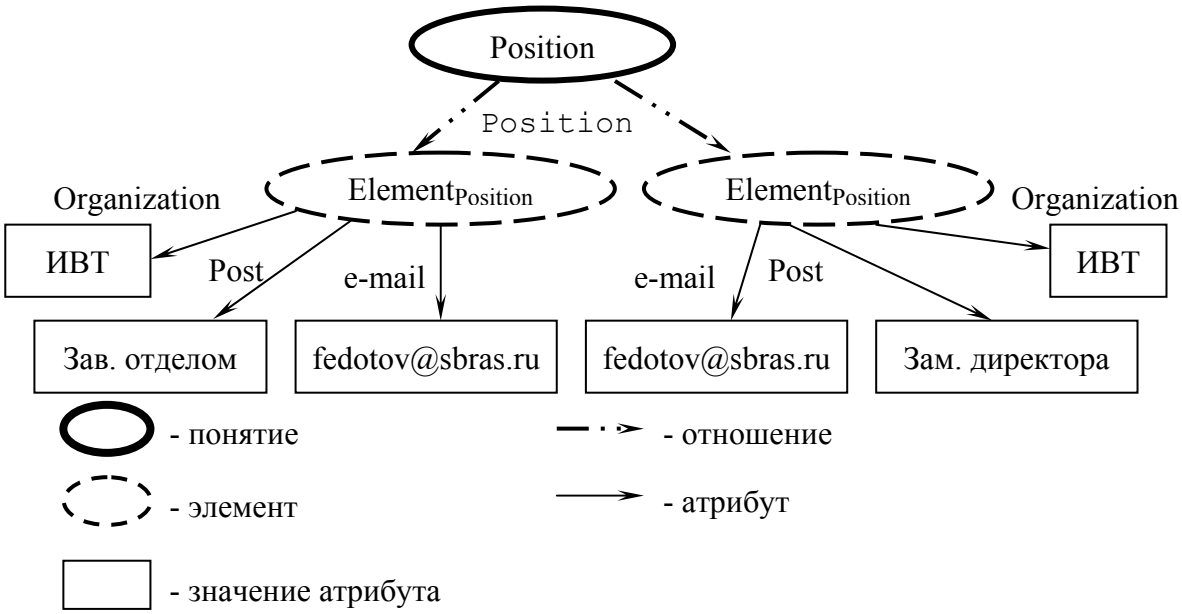

Рис.5. RDF-представление дополнительных атрибутов отношений второго типа с учетом множественного типа

должности, ключевые слова, дата назначения, дата освобождения от должности, видимость и др.

Для отношения A(R,V) аргумент R будем называть *головным документом*, а аргумент V – *подчиненным документом*.

Документ в системе может быть связан с любым количеством документов. Между двумя документами могут быть заданы прямые и обратные отношения.

Прямое отношение – отношение головного документа к подчиненному ему документу, например отношение документа «визитная карточка организации» к документу, содержащему множество подразделений, множество сотрудников или список дополнительной информации. Документ из коллекции Персона или Организации может быть связан отношением с документами из коллекции дополнительной информации, например списком дополнительных сведений.

Обратное отношение – отношение подчиненного документа к головному документу.

Для односторонних отношений родительский документ всегда знает свои дочерние документы, а дочерний документ ничего не знает о своем родителе. Для обеспечения навигации по коллекциям необходим учет обратных отношений о документе.

Выделяя два вида отношений между документами, мы решаем две задачи:

- Навигация по коллекциям (рис.1) (навигационное дерево);
- установление связей между документами (гиперссылки, вставки).

Исходя из свойств отношений второго типа, в документе можно выделить два типа элементов:

1) элементы, содержание которых не зависит от значений атрибутов отношения;
2) элементы, содержание которых может зависеть от значений атрибутов отношения (например, от должности персоны в организации зависит служебная информация).

Заметим, что элементы второго типа могут содержать списки ссылок на другие документы, списки вставок.

На рис.2 изображены прямые связи коллекции «Организации» и коллекции «Персоны», на рис.3 – коллекции «Публикации» и коллекции «Персоны», на рис.4 – структура дополнительных атрибутов отношений второго типа с учетом множественного типа.

Однако использование указанной схемы не решает всех проблем, возникающих при создании информационно-справочных систем для научного сообщества, например, проблему утраты с течением времени актуальности информации, сконцентрированной вокруг организаций, сообществ и т.п. Так, для нас может представлять интерес метод Бубнова – Галеркина решения операторных уравнений или сама биография И.Г.Бубнова, но вряд ли мы будем искать эту информацию посредством поиска сведений о Морской академии или Опытовом судостроительном бассейне, где служил Бубнов.

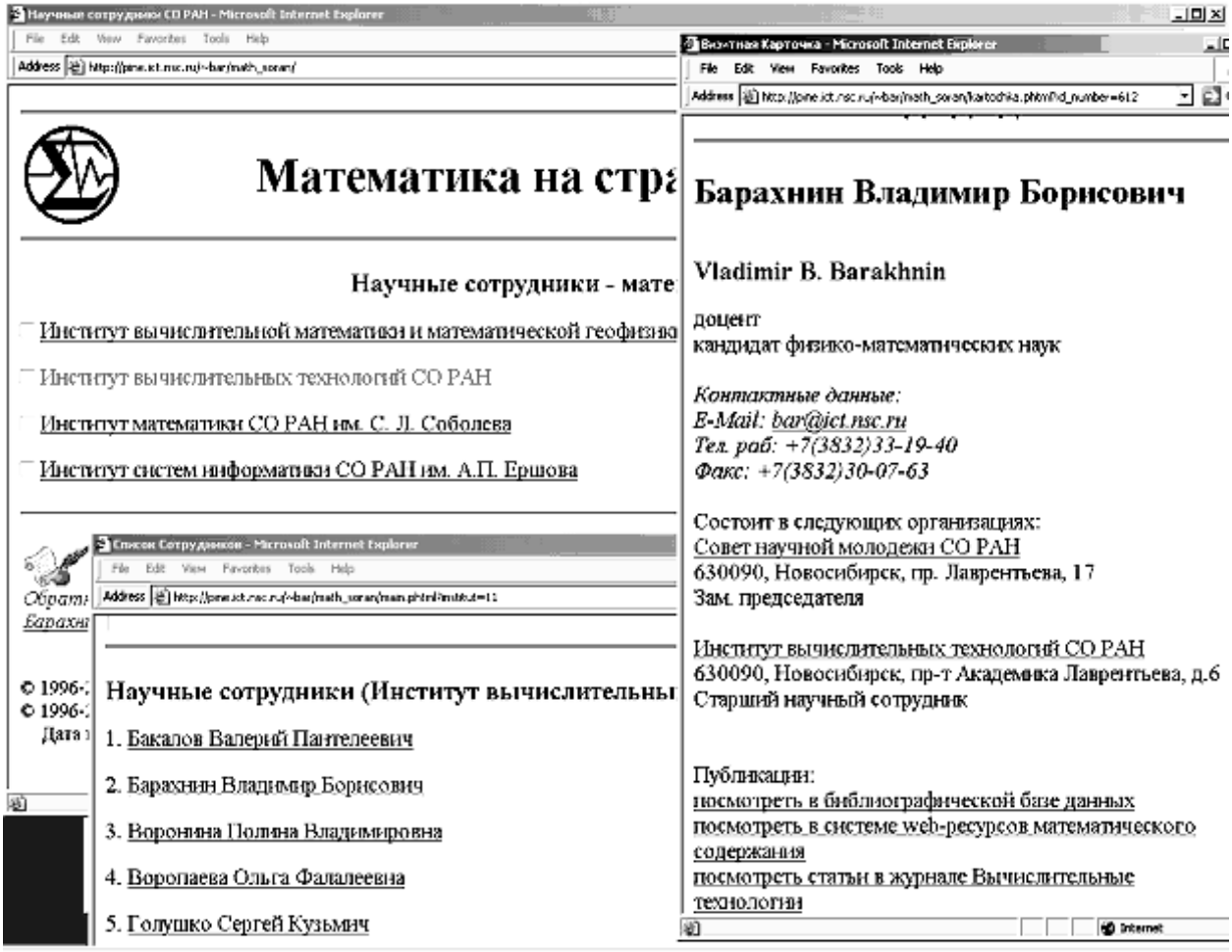

Рис.6. Информационная система «Математики СО РАН»

Поэтому информацию целесообразно группировать следующим образом:

1) вокруг персон ученых;
2) вокруг понятий и фактов науки;
3) вокруг описываемых наукой реалий окружающего мира (для естественных и некоторых гуманитарных наук).

В рамках данной статьи нас интересует первый подход (методика построения информационных моделей в рамках подхода с использованием тезаурусов изложена, в частности, в [11]; третий подход опирается, главным образом на систематику предметов, изучаемых в рамках соответствующей науки).

Для информационной модели системы, в основе которой находится персона, при поиске возникает необходимость сопоставить персоне все ее позиции (в том числе и относительно публикаций), т.е. пользоваться подходом, обратным описанному выше. Решение этой задачи с помощью контекстных запросов (даже к конкретному полю) не всегда удобно, т.к. может привести к выдаче нерелевантных документов. Таким образом, возникает потребность в построении обратной модели отношений, которая носила бы достаточно универсальный характер.

Таким образом, документы информационной системы сгруппированы по следующему принципу: имеется специально выделенная коллекция “Персоны” и множество других коллекций: “Публикации”, “Организации”, “Сообщества” (т.е. советы, общества, журналы и др.), причем все отношения строятся вокруг персон.

Персона может занимать различные позиции: быть автором или редактором публикации, занимать некоторую должность в организации, быть председателем или членом совета и т.д. Все эти случаи представляются одним типом отношения `Position`, который может принимать различные наименования (*директор*, *аспирант*, *председатель совета*, *автор* и т.д.)

На рис. 5 приведена RDF-схема, описывающая представление множественного элемента `Position` из схемы данных коллекции персон, содержащего должность персоны. Персона *А.М.Федотов* связана с организацией *ИВТ СО РАН* отношениями «зам. директора по науке» и «зав. отделом».

Исходя из документов, связанных отношениями, мы можем по запросу генерировать следующее внутреннее представление документа:

```
<id = "14">
<LastName>Федотов</LastName>
<FirstName>Анатолий</FirstName>
<MiddleName>Михайлович</MiddleName>
<Position>
  <organization name="ИВТ СО РАН">
     <post>Зав. отделом</post>
     <post>Зам. директора по науке</post>
  </organization>
</Position>
```

Использование модели в информационной системе "Web-ресурсы математического содержания"

Созданная в ИВТ СО РАН информационная система "Web-ресурсы математического содержания" [12, 13] предназначена для каталогизации математических интернет-ресурсов с целью обеспечения релевантного поиска нужной информации. В процессе работы с системой выяснилось, что древовидной структуры информации, упорядочивающей документы по их типу (персона, общество, институт, отдел, лаборатория, группа, факультет, кафедра, научная школа, конференция, семинар, издательство, журнал, книга, статья, проект, пакет программ, библиотека, коллекция, база данных, форум), а также в соответствии с "Классификатором математических сущностей", используемым Американским и Европейским математическими обществами, явно недостаточно. В новой версии системы предусмотрено установление внутренних связей между документами в соответствии с описанной выше методикой. Эти связи дают возможность при выводе информации об организации так или иначе отображать информацию о персонах, в ней работающих, и о публикациях этих персон, при выводе информации о персоне давать гиперссылку на сайт соответствующей организации и список ресурсов – публикаций данной персоны и т.д. На первом этапе указанная модификация затрагивает ресурсы раздела "Математики СО РАН" [14]).

Информацию раздела можно разделить на две части: биографическую и библиографическую. В основе биографической части лежит информационная система "База данных организаций и сотрудников СО РАН" [15]: списки организаций и сотрудников, а также сведения о связях между элементами названных списков и атрибуты этих связей (должность, контактная информация и т.п.). Указанная информация отображается в визитной карточке персоны (см. рис.6 Информационная система «Математики СО РАН»), причем, если персона занимает несколько должностей (в одной или разных организациях), то отображается информация обо всех должностях.

Библиографическая часть состоит из разнообразных баз данных: публикации сотрудников того или иного института, содержание издаваемых в СО РАН журналов, собственная библиографическая база данных системы "Web-ресурсы математического содержания" и т.п. В настоящее время информация из различных баз данных представляется независимо, однако планируется создать единую систему вывода, устраняющую дубликаты.

Литература

[1] Портал MathTree. [http://www.mathtree.ru]

[2] Портал Math-Net.RU. [http://www.math-net.ru]

[3] Портал MacTutor History of Mathematics. [http://www-history.mcs.st-and.ac.uk/ history/]

[4] *Жижимов О.Л.*, *Мазов Н.А.* Принципы построения распределенных информационных систем на основе протокола Z39.50. Новосибирск: Изд-во ИВТ СО РАН, 2004. - 361 с.

[5] *Валиев М.К, Китаев Е.Л., Слепенков М.И.* Использование службы директорий LDAP для представления метаинформации в глобальных вычислительных системах. [http://www.keldysh.ru/metacomputing/ism99.html]

[6] *Атре Ш.* Структурный подход к организации баз данных / Пер.с англ.-М.:Финансы и статистика,1983.-317 с.

[7] *Ульман Дж.* Основы систем баз данных / Пер.с англ.-М.:Финансы и статистика,1983.-334 с.

[8] *Мейер Д.* Теория реляционных баз данных / Пер.с англ.-М.:Мир,1987.-608 с.

[9] Концепция открытых систем // Материалы к межотраслевой Программе "Развитие и применение открытых систем".

[10] Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation 22 February 1999 [http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/]

[11] *Барахнин В.Б.* Разработка тезауруса предметной области "Математика" // Вычислительные технологии, т. 8, Региональный вестник Востока, N 3 (19), совместный выпуск. - 2003. - Часть 1. - С. 111-115.

[12] *Барахнин В.Б., Гуськов А.Е., Клименко О.А., Рычкова Е.В., Столяров С.В.* Информационная система "Web-ресурсы математического содержания" // Материалы конференции молодых ученых, посвященной М.А.Лаврентьеву. Новосибирск, 17-19 ноября, 2004. Часть I. - 2004.- С. 23-27.

[13] Информационная система «Математики СО РАН» [http://www.sbras.ru/sbras/math_soran/]
[14] Информационная система «Web-ресурсы математического содержания» [http://www-sbras.nsc.ru/win/elbib/data/show_page.dhtml?2+184]
[15] *Шокин Ю.И., Федотов А.М., Клименко О.А., Леонова Ю.В.* Содержательное наполнение справочно-информационной системы научного сообщества // Вычислительные технологии. (Совместный выпуск). Вестник КазНУ им. аль-Фараби. Серия: Математика, механика, информатика. Ч. 4. - 2004. - Т. 42. - № 3. - С. 346-350.

Application of model of directed relations between documents for construction of information systems of scientific community

Barakhnin V.B., Leonova J.V.

In this paper the information model of horizontal directed relations between documents is considered on the basis of the binary relations with additional attributes. Its application for construction of scientific information systems is considered.