ПОЛИДОМЕННЫЕ МОДЕЛИ В СИСТЕМАХ ОЦЕНКИ ИННОВАЦИОННОГО ПОТЕНЦИАЛА И РЕЗУЛЬТАТИВНОСТИ НАУЧНЫХ ИССЛЕДОВАНИЙ[1)]

POLYDOMAIN MODELS FOR EVALUATION SYSTEMS OF INNOVATIVE POTENTIAL AND PERFORMANCE OF RESEARCHES

И.М. Зацман

ИПИ РАН, Москва

Рассматриваются модели интеллектуальных автоматизированных систем, предназначенных для мониторинга и оценки инновационного потенциала и результативности направлений научных исследований. Рассматриваемые модели являются сочетанием лексико-семантического, информационного, алгоритмического, математического и ряда других компонентов.

Введение

В докладе вводится понятие **полидоменных моделей**, включающих лексико-семантический, информационный, алгоритмический, математический, биоинформационный, геоинформационный и другие компоненты. Подобные сочетания перечисленных компонентов предназначены, в первую очередь, для моделирования и проектирования интеллектуальных автоматизированных систем в слабоформализуемых и институционально сложных сферах применения, например, в сфере науки, для мониторинга правового пространства и правоприменительной практики, в сфере инноваций, для глобального мониторинга и оценки последствий выбросов парниковых газов, в патентной сфере или для мониторинга и прогнозирования чрезвычайных ситуаций.

Полидоменные модели являются развитием логико-лингвистических моделей управления [2], ориентированным на моделирование сложных институциональных систем от макроуровня до наноуровня [3].

Предлагаемый подход к интеграции перечисленных компонентов в рамках единой модели позволяет многоаспектно структурировать и описывать сочетания абстрактных и конкретных знаковых и формально-символьных образований, например, математические и химические формулы, электронные карты, корпуса текстов на естественных языках, потоковые аудио- и видеообъекты, вербально-образные тезаурусы и онтологии, а также эксплицировать типологию отношений между ними.

В докладе цель разработки и назначение полидоменных моделей рассматриваются на примере систем верифицируемого мониторинга и оценки инновационного потенциала и результативности направлений научных исследований. Сложность создания подобных систем заключается в том, что они предназначены для применения одновременно в двух слабоформализуемых и институционально сложных сферах применения: в сфере науки и в сфере инноваций. Актуальность создания подобных систем вызвана тем, что наиболее существенным препятствием на пути применения методов программно-целевого планирования и финансирования в сфере науки является нерешенность проблемы мониторинга, анализа и **верифицируемой** индикаторной оценки инновационного потенциала и результативности субъектов сферы науки, направлений, программ и проектов научных исследований (далее – проблема верифицируемого мониторинга).

За последние несколько десятилетий сформировался спектр индикаторов, включая количество опубликованных по проекту научных статей, их цитируемость, импакт-факторы журналов, в которых опубликованы статьи и т.д. Однако использование подобных индикаторов в процессах планирования и финансирования при оценке направлений, программ и проектов научных исследований должно позволять решать одновременно и задачи верификации используемых значений индикаторов и сопоставления значений одних и тех же индикаторов, полученных с помощью разных систем мониторинга и оценки инновационного потенциала и результативности направлений научных исследований (далее – системы верифицируемого мониторинга или СВМ).

В настоящее время изменились роль мониторинга в сфере науки. Данные мониторинга и определение с их помощью значений индикаторов до настоящего времени практически не влияли на бюджетный процесс и

[1)] Работа выполнена при частичной поддержке РГНФ, проект № 06-02-04043а

[2]. Поспелов Д.А. Логико-лингвистические мо-дели в системах управления.- М.: Энергоиздат, 1981.

[3]. Клейнер Г.Б. Эволюция институциональных систем.- М.: Наука, 2004.

распределение бюджетных средств в сфере науки. Однако сейчас планируется, что через нескольких лет все 100% бюджета страны, включая научный бюджет, будут распределяться по целевым программам с использованием индикаторов [4, 5]. Это делает проблему верифицируемого мониторинга и проектирование СВМ, включая решение задач верификации используемых значений индикаторов и сопоставления их значений, весьма актуальными.

Таким образом, при проектировании СВМ, необходимо предусмотреть выполнение ряда условий, соблюдение которых позволит верифицировать и сопоставлять значения индикаторов в процессе планирования и распределения бюджетных средств.

Пример традиционного моделирования

Рассмотрим пример широкого используемого в сфере науки индикатора – возрастная структура исследователей. В примере на рис. 1 две полужирные кривые относятся к заявителям РФФИ, местом работы которых является РАН, и две тонкие кривые относятся к остальным заявителям. Серым цветом обозначены кривые, построенные с использованием результатов обработки данных мониторинга за 1997 год, а черным – за 2002 год.

По оси абсцисс указан возраст заявителей, определенных на основе формы 2 из заявок РФФИ. На этой оси в явном виде возраст заявителей указан с шагом в пять лет от 16 лет до 101 года, но вычисления выполнены с шагом в один год. По оси ординат указан процент заявителей каждого возраста с шагом в один год от 16 лет до 101 года. Изображенные непрерывные кривые являются интерполяцией вычисленных дискретных функций.

На этом рисунке видно, что арифметическое отношение долей молодых заявителей РАН в 2002 и 1997 годах превышает единицу. Численные отношения долей заявителей РАН, выраженные в процентах, приведены в нижней строке табл. 1 с шагом в пять лет. Рис. 1 и табл. 1 взяты из [6].

Приведенные кривые и таблица отражают старение исследователей РАН и других ведомств, так как острота пика молодых исследователей со временем увеличивается и после пика черные кривые уходят резко вниз.

Если использовать математические обозначения, то значения этого индикатора можно представить как векторную функцию $\mathbf{f}=(f_1, f_2)$, где f_1 является индикатором для заявителей РАН (см. на рис. 1 серую и черную полужирные кривые для 1997 и 2002 годов, соответственно), а f_2 – индикатором для всех остальных заявителей (серая и черная тонкие кривые для 1997 и 2002 годов, соответственно).

Векторная функция **f** зависит от дискретной переменной возраста заявителей t_i, где $t_0=16$, $t_1=17$ и так далее до 101 года и дискретной переменной T_j, где $T_0=1997$, $T_1=2002$. На рис. 1 приведены две серые кривые для $T_0=1997$ и две черные кривые для $T_1=2002$. В краткой математической форме этот индикатор возрастной структуры исследователей имеет вид:

f (t_i, T_j), где t_i и T_j являются дискретными переменными, i, j=0,1, … ; $\mathbf{f}=(f_1, f_2)$.

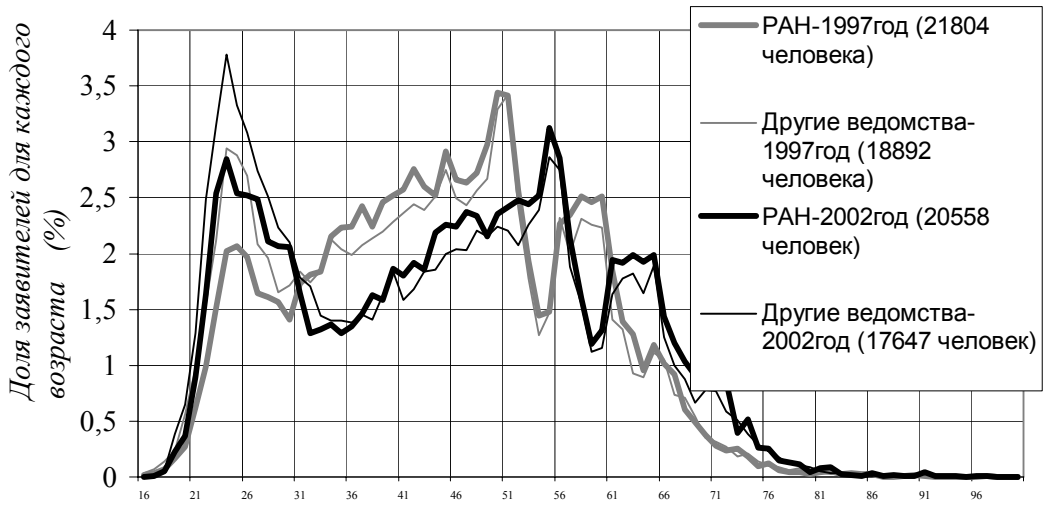

Рис. 1. Возрастная структура заявителей на получение грантов РФФИ в 1997 и 2002 годах

[4]. Стенограмма выступления Заместителя Председателя Правительства РФ А.Д. Жукова на VI Международной конференции "Модернизация экономики и выращивание институтов"; http:// www.hse.ru/temp/2005/files/04_06_2005_jukov.doc.

[5]. Зацман И.М Терминологический анализ нормативно-правового обеспечения создания систем мониторинга в сфере науки // Экономическая наука современной России. № 4, 2005. С. 114-129.

[6]. Зацман И.М. Информационные ресурсы для систем мониторинга в сфере науки // Системы и средства информатики. Вып. 15.- М.: Наука, 2005. С. 288-318.

Возраст заявителей РАН	20	25	30	35	40	45	50	55	60	65	70	75
Доля в 2002г.	0,37 %	2,54%	2,06%	1,28%	1,86%	2,25%	2,35%	3,12%	1,31%	1,98%	0,93%	0,26%
Доля в 1997г.	0,27 %	2,06%	1,41%	2,23%	2,51%	2,91%	3,43%	1,48%	2,51%	1,18%	0,37%	0,1%
Отношение долей	1,37	1,23	1,46	0,57	0,74	0,77	0,69	2,11	0,52	1,68	2,51	2,6

Таблица 1. Арифметическое отношение долей заявителей РАН в 2002 и 1997 годах

Приведенная формула отражает конечные результаты вычислений, но ничего не говорит об источнике использованных в процессе вычислений информационных ресурсах системы мониторинга, что является необходимых условием решения проблемы верифицируемого мониторинга.

Информационный компонент модели

В процессе построений кривых на рис. 1 были использованы данные о датах рождения 21804 заявителей РАН в 1997 году и 20558 заявителей РАН в 2002 году, а также даты рождения исследователей других ведомств, которые хранятся в базе персональных данных. На рис. 2 показан фрагмент схемы записи базы персональных данных [7]. На этом рисунке видно, что дата рождения (name="BirthDay") имеет три атрибута, включая год рождения (attribute name="Year"). Полужирным цветом на схеме выделены используемые далее элементы схемы (поля) и атрибуты элементов.

Для отражения схемы записи базы персональных данных в модели системы верифицируемого мониторинга добавим в число параметров векторной функции **f** (t_i , T_j) гипертекстовую ссылку http://…person-xsd.xml на место хранения этой схемы. После добавления ссылки последовательность символов **f** (t_i , T_j , http://…person-xsd.xml) перестает быть математическим объектом, так как последний параметр в скобках является информационным и после добавления ссылки мы получаем информационно-математический компонент.

```
<?xml version="1.0"?>
<xs:schema id="Persons">
              …
              <xs:element name="Sex" type="xs:string" minOccurs="0" msdata:Ordinal="1" />
              <xs:element name="Name" minOccurs="0" maxOccurs="unbounded">
                <xs:complexType>
                  <xs:attribute name="Last" form="unqualified" type="xs:string" />
                  <xs:attribute name="First" form="unqualified" type="xs:string" />
                  <xs:attribute name="Middle" form="unqualified" type="xs:string" />
                </xs:complexType>
              </xs:element>
              <xs:element name="BirthDay" nillable="true">
                <xs:complexType>
                  <xs:simpleContent msdata:ColumnName="BirthDay_Text" msdata:Ordinal="3">
                    <xs:extension base="xs:string">
                      <xs:attribute name="Day" form="unqualified" type="xs:string" />
                      <xs:attribute name="Month" form="unqualified" type="xs:string" />
                      <xs:attribute name="Year" form="unqualified" type="xs:string" />
                    </xs:extension>
                  </xs:simpleContent>
                </xs:complexType>
              </xs:element>
              …
              <xs:element name="Work" minOccurs="0" maxOccurs="unbounded">
                <xs:complexType>
                  <xs:attribute name="OrganizationId" type="xs:string" />
                  <xs:attribute name="Post" form="unqualified" type="xs:string" />
                </xs:complexType>
              </xs:element>
              …
</xs:schema>
```

Рис. 2. Схема записи базы персональных данных в сокращенном виде

[7]. Шубников С.К. Формы документов в систе-мах информационного обеспечения оценки результативности научной деятельности // Системы и средства информатики. Вып. 15.- М.: Наука, 2005. С. 59-76.

Его объединение со схемой записи базы персональных данных и с векторной функцией является частным примером полидоменной модели, состоящей из трех компонентов: схемы записи на рис. 2, математической функции **f** (t_i , T_j) и информационно-математического компонента **f** (t_i , T_j , http://…person-xsd.xml).

Добавление ссылки http://…person-xsd.xml позволяет найти год рождения заявителя, что является необходимым для вычисления индикатора возрастной структуры. Однако в этой схеме отсутствуют сведения о ведомственной принадлежности организации-места работы исследователя. В схеме записи базы персональных данных на рис. 2 содержит только уникальный идентификатор организации-места работы (attribute name="OrganizationId") в базе реквизитов организаций и должность исследователя.

Следовательно, для вычисления индикатора возрастной структуры исследователей РАН необходимо знать схему записи базы реквизитов организаций, которая строится аналогично схеме на рис. 2. Поэтому схема записи базы реквизитов организаций не приводится.

Для отражения схемы записи базы реквизитов в модели системы верифицируемого мониторинга добавим в число параметров информационно-математического компонента ссылку http://…org-xsd.xml на эту схему. Это добавление является необходимым для вычисления индикатора возрастной структуры с учетом ведомственной принадлежности организации-места работы исследователя (см. рис. 1). Включение в полидоменную модель схем записи базы персональных данных и записи базы реквизитов фиксирует структуру информационных ресурсов системы мониторинга, необходимых для вычисления этого индикатора, но ничего не говорит об источнике использованных в процессе вычислений информационных ресурсах.

Однако добавление ссылок http://…persons1997.xml и http://…persons2002.xml на персональные данные исследователей, а также ссылок http://…org1997.xml и http://…org2002.xml на реквизиты организаций фиксирует ссылки на те информационные ресурсы, которые использовались при вычислении индикатора на рис. 1. Пример фрагмента записи персональных данных одного исследователя приведен на рис. 3; пример фрагмента записи реквизитов организации приведен на рис. 4 [94].

После добавления перечисленных ссылок полидоменная модель включает:

- информационный компонент в форме схем двух записей (базы персональных данных и базы реквизитов организаций);
- математический компонент в виде векторной функции **f** (t_i , T_j);
- информационно-математический компонент **f** (t_i , T_j , http://…person-xsd.xml, http://…org-xsd.xml, http://…persons1997.xml, http:// …persons2002.xml, http://…org1997.xml, http://…org2002.xml).

```
<?xml version="1.0" encoding="windows-1251"?>
<Persons>
   <Person Id="1">
         <Name Last="Иванов" First="Иван" Middle="Петрович"/>
         <Sex>1</Sex>
         <BirthDay Day="01" Month="11" Year="1977"></BirthDay>
         <Work OrganizationId="1" Post="лаб"/>
         …
   </Person>
   <Person Id="2">
         …
   </Person>
</Persons>
```

Рис. 3. Фрагмент записи персональных данных одного исследователя

```
<?xml version="1.0" encoding="windows-1251" ?>
<Organizations>
   <Organization Id="1" Name="Институт проблем информатики РАН" Ministry="РАН"
ShortName="ИПИ РАН" Tel="(495)1356260" Fax="(495)9304505" E-mail="ipiran@ipiran.ru"
PostAddressId="2" INN="" KPP="" WWW="http://www.ipiran.ru" />
         …
</Organizations>
```

Рис. 4. Фрагмент записи реквизитов организации

Алгоритмический компонент модели

Перечисленные компоненты модели содержат ссылки на использованные в процессе вычислений индикатора информационные ресурсы системы мониторинга, но ничего не говорят о том алгоритме, с помощью которого определяются значения векторной функции **f** (t_i , T_j).

Математический компонент полидоменной модели содержит следующую информацию: для вычисления **f**=(f_1, f_2) в точке (t_i , T_j) необходимо для определения f_1 разделить число исследователей РАН, подавших заявки в РФФИ в году T_j и возраст которых в этом году равнялся t_i , на общее число исследователей РАН, подавших

заявки в РФФИ в этом году (T_j), и умножить на 100 для получения процентного отношения. Для определения f_2 необходимо разделить число исследователей из других ведомств, подавших заявки в РФФИ в году T_j и возраст которых в этом году равнялся t_i, на общее число исследователей из других ведомств, подавших заявки в РФФИ в этом году (T_j), и умножить на 100.

Например, в 1997 году общее число исследователей РАН, подавших заявки в РФФИ, было равно 21804 человека. Из них 59 человек достигли 20-летнего возраста. Следовательно, f_1 (20, 1997)=59:21804*100=0,27%, т.е. доля 20-летних исследователей РАН, подавших заявки в РФФИ в 1997 году составила 0,27% от общего числа исследователей РАН, подавших заявки в РФФИ.

Однако подобное математическое описание не раскрывает алгоритма, с помощью которого было получено число 59, которое является исходным для вычисления f_1 в математическом компоненте полидоменной модели. В системе мониторинга это число определяется в результате поиска записей в базе персональных данных атрибута Year (см. рис. 3), который содержит последовательность четырех цифр «1977» при условии, что в соответствующей записи базы реквизитов организаций (т.е. для организации-места работы исследователя) атрибут Ministry (см. рис. 4) содержит последовательность букв «РАН». Число найденных записей и будет равно 59.

Следовательно, кроме математического компонента, полидоменная модель должна включать формальную запись всех тех алгоритмов (обозначим это множество алгоритмов как {A}), которые используются для вычисления индикатора возрастной структуры исследователей. Кроме того, в информационно-математический компонент необходимо добавить в качестве еще одного параметра ссылку http://…programA на исходный текст программы, реализующей множество алгоритмов {A}. Отметим также необходимость добавления множества алгоритмов верификации, которое обозначим как {B}.

После добавления этой ссылки и алгоритмов полидоменная модель включает:

- информационный компонент в форме схем двух записей (базы персональных данных и базы реквизитов организаций);
- математический компонент в виде векторной функции $\mathbf{f}(t_i, T_j)$;
- множество алгоритмов вычисления индикатора возрастной структуры {A};
- множество алгоритмов верификации результатов вычислений индикатора возрастной структуры {B};
- информационно-математический компонент $\mathbf{f}(t_i, T_j$, http://…programA, http://…programB, http://…person-xsd.xml, http://…org-xsd.xml, http://…persons1997.xml, http://…persons2002.xml, http://…org1997.xml, http://…org2002.xml).

Лексико-семантический компонент модели

Перечисленные компоненты полидоменной модели включают все необходимое для вычисления и верификации результатов вычисления индикатора возрастной структуры исследователей. Поэтому возникает естественный вопрос о необходимости лексико-семантического компонента.

Рассматриваемые в докладе полидоменные модели являются средством описания и инструментом проектирования систем верифицируемого мониторинга, которые служат для определения широкого спектра индикаторов, который обозначим как множество векторных функций {**f**}, а не только одного индикатора возрастной структуры. Среди них есть индикаторы, для определения значений которых необходимо разрабатывать лексико-семантический компонент полидоменной модели.

Рассмотрим в качестве примера индикатор инновационного потенциала теоретических исследований по информатике, проводимых в РАН. Название этого индикатора практически ничего не говорит разработчику системы верифицируемого мониторинга о его смысле. Следовательно, до начала разработки системы необходимо иметь эксплицитно выраженную концептуализацию этого индикатора. Отметим, что концептуализация и экспликация смысла индикаторов являются одной из основных задач при разработке лексико-семантического компонента полидоменной модели.

Остановимся на индикаторе инновационного потенциала подробнее. Сложность концептуализации этого индикатора заключается в том, что он относится одновременно к сфере фундаментальных исследований, в которой используются научные классификаторы областей знаний и научных специальностей (например, классификатор РФФИ), и к сфере инноваций, в которой используется Международная патентная классификация (МПК). В настоящее время отсутствует таблицы перехода от рубрик научных классификаторов к МПК. Поэтому недостаточно зарубрицировать теоретические исследования по информатике с помощью рубрик научного классификатора. Необходимо дополнительно составить лексико-семантический портрет каждой рубрики, например, с помощью дескрипторов тезауруса и отношений между ними. В результате сопоставления лексико-семантических портретов рубрик и корпуса описания изобретений, имеющих индексы МПК, строится таблицы перехода от рубрик научных классификаторов к МПК для теоретических исследований по информатике.

В приведенном примере лексико-семантический компонент полидоменной модели должен включать научный классификатор, МПК и тезаурус, создание и ведение которого для предметной области информатики является трудоемкой задачей, а также таблицу перехода от рубрик научных классификаторов к МПК.

Тогда индикатор инновационного потенциала теоретических исследований по информатике, проводимых в РАН можно определить как отношение научных статей исследователей РАН по информатике, цитируемых в

массиве описаний изобретений, имеющих индексы МПК, отраженные в таблице перехода, к общему числу научных статей, цитируемых в этом массиве описаний изобретений. Приведенный пример наглядно иллюстрирует сферу применения лексико-семантического компонента полидоменной модели в процессе разработки систем верифицируемого мониторинга.

Заключение

В работе рассмотрены несколько частных случаев наборов компонентов полидоменной модели. В общем случае этот класс моделей интеллектуальных автоматизированных систем (ИАС) включает:

- информационный компонент, например в форме набора версионных схем информационных ресурсов ИАС, т.е. в модели ослеживается изменение схем в зависимости от времени эксплуатации ИАС;
- математический компонент в виде множества векторных функций {**f**} или целевых функционалов;
- множество прикладных алгоритмов {A}, необходимых для вычисления векторных функций {**f**} или целевых функционалов;
- множество алгоритмов {B} для верификации результатов вычислений векторных функций {**f**} или целевых функционалов;
- коммуникационный компонент модели (в частном случае, информационно-математический компонент);
- лексико-семантический компонент, например, версионные классификаторы, тезаурусы, системы онтологий, т.е. в модели ослеживается их изменение в зависимости от времени;
- тематические компоненты, например, био- или геоинформационные модели;
- описание семиотической знаковой системы ИАС.

Необходимость включения последнего компонента мотивирова тем, что современные компьютерные системы основаны на кодировании знаковых примитивов. Такое кодирование ориентировано на те формы знаков, которые могут быть представлены в виде конкатенаций знаковых примитивов, например, линейных текстов.

Для вербально-образных и других сложных текстов [[1], [2]] иногда необходимо в явной форме эксплицировать знаковую систему ИАС, что и должно обеспечиваться описанием семиотической знаковой системы ИАС.

Список литературы

.

1. Поспелов Д.А. Логико-лингвистические мо-дели в системах управления.- М.: Энергоиздат, 1981.
2. . Клейнер Г.Б. Эволюция институциональных систем.- М.: Наука, 2004.
3. . Стенограмма выступления Заместителя Председателя Правительства РФ А.Д. Жукова на VI Международной конференции "Модернизация экономики и выращивание институтов"; http:// www.hse.ru/temp/2005/files/04_06_2005_jukov.doc.
4. . Зацман И.М Терминологический анализ нормативно-правового обеспечения создания систем мониторинга в сфере науки // Экономическая наука современной России. № 4, 2005. С. 114-129.
5. . Зацман И.М. Информационные ресурсы для систем мониторинга в сфере науки // Системы и средства информатики. Вып. 15.- М.: Наука, 2005. С. 288-318.
6. . Шубников С.К. Формы документов в систе-мах информационного обеспечения оценки результативности научной деятельности // Системы и средства информатики. Вып. 15.- М.: Наука, 2005. С. 59-76.
7. . Зацман И.М. Концептуальный поиск и качество информации.– М.: Наука, 2003. 271 с.
8. . Зацман И.М. Семантическое, информацион-ное и знаковое кодирование патентных документов электронных библиотек // Труды Седьмой Всероссийской конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (Ярославль, 4-6 октября 2005г.).– Ярославль: Ярославский госуниверситет, 2005. С. 112-121.

1. Зацман И.М. Концептуальный поиск и качество информации.– М.: Наука, 2003. 271 с.
2. Зацман И.М. Семантическое, информацион-ное и знаковое кодирование патентных документов электронных библиотек // Труды Седьмой Всероссийской конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (Ярославль, 4-6 октября 2005г.).– Ярославль: Ярославский госуниверситет, 2005. С. 112-121.