ФЕДЕРАЛЬНОЕ АГЕНТСТВО НАУЧНЫХ ОРГАНИЗАЦИЙ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ
ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 5 Москва 2018

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 004.22:004.8

Д.В. Виноградов

О представлении объектов битовыми строками для ВКФ-метода

Предложен и исследован алгоритм построения кодирования битовыми строками объектов, описываемых дискретными признаками, на значениях каждого из которых задана бинарная операция сходства. При этом кодировании операция побитового умножения должна соответствовать покомпонентной операции сходства.

Ключевые слова: *битовые строки, операция сходства, супремум-плотное подмножество, супремум-неразложимые элементы*

ВВЕДЕНИЕ

Вероятностный метод машинного обучения, основанный на бинарной операции сходства, был предложен автором работы [1]. При создании этого подхода использовались идеи ДСМ-метода автоматического порождения гипотез [2] и анализа формальных понятий (АФП) [3].

Ключевым моментом, обеспечивающим высокую вычислительную эффективность предложенного подхода, является представление обучающих и тестовых примеров в виде битовых строк так, чтобы операция сходства соответствовала побитовому умножению. То, что такое представления всегда существует, было установлено проф. Рудольфом Вилле в теореме, которая получила название Фундаментальная теорема АФП. Современное доказательство этой теоремы приводится в книге [4].

В настоящей работе мы предложим алгоритм вычисления такого кодирования объектов битовыми строками, которое имеет кратчайшую длину. Основная цель настоящей работы – доказать корректность этого алгоритма. Здесь мы существенно опираемся на идеи АФП, которые были развиты для современного доказательства теоремы Р.Вилле.

Для апробации ВКФ-метода нами была создана программная система, названная ВКФ-системой, которая была с успехом применена к нескольким массивам из репозитория данных для тестирования алгоритмов машинного обучения. Но при этом кодирование было ручным. С помощью изложенного алгоритма и этот этап ВКФ-исследования становится автоматизированным.

ВСПОМОГАТЕЛЬНЫЕ ОПРЕДЕЛЕНИЯ И ФАКТЫ

Сходство является бинарной операцией на множестве X, объемлющем множество объектов, т. е. представляет собой отображение $\cap: X \times X \to X$. Операция сходства должна удовлетворять аксиомам *нижней полурешетки*:

$$x \cap x = x \tag{1}$$

$$x \cap y = y \cap x \tag{2}$$

$$x \cap (y \cap z) = (x \cap y) \cap z \tag{3}$$

Для выражения тривиальности сходства имеется специальный *пустой фрагмент* $\varnothing$ со свойством наименьшего элемента.

$$x \cap \varnothing = \varnothing \tag{4}$$

Во многих прикладных задачах обучающие и тестовые примеры описываются набором признаков. Номинальные и порядковые признаки (с конечным числом значений) легко превращаются в нижнюю полурешетку. Если признак измеряется в шкале разностей или отношений, то приходится использовать группировку значений. Но в общем случае полурешетка является более общим понятием, так как всегда определяет порядок:

$$x \le y \equiv (x \cap y = x) \tag{5}$$

Важнейшим примером для нас будет нижняя полурешетка, состоящая из битовых строк фиксированной длины с побитовым умножением в качестве операции сходства. Пустым фрагментом будет являться строка, состоящая из одних нулей. Каждый бит может быть отождествлен с бинарным признаком, тогда битовая строка соответствует множеству признаков, в которых встречаются единицы. При этом операция сходства соответствует пересечению множеств признаков, а пустой фрагмент – пустому множеству признаков.

Важность этого примера объясняется тем, что операция побитового умножения допускает эффективную реализацию на современных ЭВМ. Существуют специальные классы объектов (например, *dynamic_bitset* в C++), реализующие удобное оперирование битовыми строками. Более того, компиляторы допускают удобное векторное распараллеливание такого кодирования.

Однако теорема Рудольфа Вилле утверждает, что эта конструкция позволяет построить любую конечную решетку из нижней полурешетки битовых строк с операцией побитового умножения, если к ней добавить наибольший элемент.

Собирая вместе битовые строки, представляющие объекты, мы получаем прямоугольную таблицу I, которую будем называть *формальным контекстом* [2]. Формальный контекст можно понимать как бинарное отношение между элементами множества O, которые называем *именами объектов* (или даже объектами), и элементами множества F, которые называем *признаками*. Если в строке, соответствующей объекту $o \in O$, и столбце, соответствующем фрагменту $f \in F$, стоит единица, то мы говорим, что *объект o обладает признаком f*, и обозначаем это через *oIf*. В противном случае, говорим, что *объект o не имеет признака f*.

Для подмножества объектов $A \subseteq O$ его *сходством* называется подмножество

$A' = \{f \in F : \forall o \in A[olf]\} \subseteq F$. Полагаем $\varnothing' = F$.

На самом деле, это определение совпадает с последовательным вычислением побитового умножения строк, соответствующих отобранным во множество A объектов.

Для подмножества $B \subseteq F$ признаков его *сходством* называется подмножество

$B' = \{o \in O : \forall f \in B[olf]\} \subseteq O$. Полагаем $\varnothing' = O$.

Сформулируем две простые леммы, прямо выводящиеся из определения, которые будут применяться в последующем изложении.

Лемма 1. ***Для*** $A_1 \subseteq O$ ***и*** $A_2 \subseteq O$ ***выполняется*** $(A_1 \cup A_2)' = A_1' \cap A_2'$ ***. Для*** $B_1 \subseteq F$ ***и*** $B_2 \subseteq F$ ***выполняется*** $(B_1 \cup B_2)' = B_1' \cap B_2'$**.**

Лемма 2. ***Операции сходства удовлетворяют следующим условиям (задают соответствие Галуа)****:*

$$\forall A \subseteq O[A \subseteq A''] \quad \forall B \subseteq F[F \subseteq F''] \tag{6}$$

$$\forall A_1 \forall A_2 [A_1 \subseteq A_2 \Rightarrow A_1' \supseteq A_2']$$
$$\forall B_1 \forall B_2 [B_1 \subseteq B_2 \Rightarrow B_1' \supseteq B_2'] \tag{7}$$

$$\forall A \subseteq O[A' \subseteq A'''] \quad \forall B \subseteq F[B' \subseteq B'''] \tag{8}$$

Определение 1. Пару $\langle A, B \rangle$ назовем ***ВКФ-кандидатом***, если $A = B' \subseteq O$ и $B = A' \subseteq F$.

В анализе формальных понятий такие пары называют *формальными понятиями*, но мы предпочитаем сменить название, так как оригинальное название подвергается обоснованной критике со стороны специалистов по философии и искусственному интеллекту.

Определение 2. ***Порядок*** на ВКФ-кандидатах зададим правилом $\langle A_1, B_1 \rangle \le \langle A_2, B_2 \rangle$, если $B_1 \subseteq B_2$.

В анализе формальных понятий определение порядка задается двойственным образом. Наше определение соответствует традиции отечественной школы.

Прямая часть теоремы Рудольфа Вилле составляет утверждение следующей теоремы, доказательство которой легко следует из определений:

Теорема 1. *Для произвольного формального контекста* $I \subseteq O \subseteq F$ *множество всех ВКФ-кандидатов образует решетку* **L** *относительно операций:*

$$\langle A_1, B_1\rangle \cap \langle A_2, B_2\rangle = \left\langle (A_1 \cup A_2)'', B_1 \cap B_2 \right\rangle \qquad (9)$$

$$\langle A_1, B_1\rangle \cup \langle A_2, B_2\rangle = \left\langle A_1 \cap A_2, (B_1 \cup B_2)'' \right\rangle \qquad (10)$$

Введем ключевое понятие для последующего изложения:

Определение 3. Подмножество элементов $S \subseteq \boldsymbol{L}$ решетки $\boldsymbol{L}$ называется $\cap$**-плотным**, если для любого элемента $x \in \boldsymbol{L}$ найдется такое подмножество $X \subseteq S$, что $x = \cap X$. Подмножество элементов $S \subseteq \boldsymbol{L}$ решетки $\boldsymbol{L}$ называется $\cup$**-плотным**, если для любого элемента $x \in \boldsymbol{L}$ найдется такое подмножество $X \subseteq S$, что $x = \cup X$.

Очевидно, что все $\boldsymbol{L}$ является как $\cap$-, так и $\cup$-плотным. Более полезный пример доставляется следующим утверждением:

Лемма 3. ***Для решетки ВКФ-кандидатов L, порождаемой формальным контекстом*** $I \subseteq O \times F$ ***образ*** $g(O) = \{g(o) : o \in O\}$ ***отображения*** $g : O \to \boldsymbol{L}$, ***задаваемого правилом*** $g(o) = \left\langle \{o\}'', \{o\}' \right\rangle$, ***является*** $\cap$***-плотным подмножеством, а образ*** $h(F) = \{h(f) : f \in F\}$ ***отображения*** $h : F \to \boldsymbol{L}$, ***задаваемого правилом*** $h(f) = \left\langle \{f\}', \{f\}'' \right\rangle$, ***является*** $\cup$***-плотным подмножеством.***

Следующая теорема составляет содержание обратной части теоремы Рудольфа Вилле:

Теорема 2. *Пусть для любой конечной решетки* $\boldsymbol{L}$ *найдутся такие два множества O и F с отображениями* $g : O \to \boldsymbol{L}$ *и* $h : F \to \boldsymbol{L}$, *что* $g(O) \subseteq \boldsymbol{L}$ – $\cap$*-плотное подмножество, а* $h(F) \subseteq \boldsymbol{L}$ – $\cup$*-плотное подмножество. Тогда, полагая* $olf \Leftrightarrow g(o) \geq h(f)$, *мы получим формальный контекст, решетка ВКФ-кандидатов которого будет изоморфна исходной решетке* $\boldsymbol{L}$.

Ясно, что тождественные отображения $g = id : \boldsymbol{L} \to \boldsymbol{L}$ и $h = id : \boldsymbol{L} \to \boldsymbol{L}$ удовлетворяют условию Теоремы 2. Именно так и проходило первоначальное доказательство Рудольфа Вилле.

К сожалению, этот вариант обычно порождает слишком большой формальный контекст. Множество объектов обычно бывает задано извне. Чтобы выбрать минимальное подмножество $F \subseteq \boldsymbol{L}$, нам необходимо ввести понятие $\cup$-неразложимых элементов.

Определение 4. Элемент $x \in \boldsymbol{L}$ решетки $\boldsymbol{L}$ назовем $\cup$**-неразложимым**, если $x \neq \varnothing$ и для любых $y, z \in \boldsymbol{L}$ если $y < x$ и $z < x$, то $y \cup z < x$.

Простые вычисления доказывают следующий результат:

Лемма 4. ***Для любой конечной решетки L любое надмножество всех*** $\cup$***-неразложимых элементов образует*** $\cup$***-плотное подмножество.***

АЛГОРИТМ КОДИРОВАНИЯ И ЕГО СВОЙСТВА

Если предполагать, что объекты описываются признаками, имеющими структуру нижней полурешетки, а операция сходства вычисляется покомпонентно, то достаточно найти в полурешетке для каждого признака $\cup$-неразложимые элементы, организовать кодирование битовыми строками, а затем все такие битовые подстроки соединить вместе в единую битовую строку.

Сосредоточимся на кодировании множества V значений одного признака. Так как операция сходства порождает порядок (Уравнение (5)), то отношение накрытия ($x < y \equiv x < y \,\&\, \neg\exists z[x < z < y]$) задает ориентированный ациклический граф.

Предлагаемый алгоритм кодирования состоит из четырех частей. Сначала осуществляется топологическая сортировка.

Определение 5. Линейный порядок $V[0] < V[1] < \ldots < \ldots < V[n-1]$ назовем **топологической сортировкой**, если $\forall i, j[V[i] < V[j] \Rightarrow i < j]$.

Известно, что топологическую сортировку можно выполнить за $O(|V| + |V|^2)$ шагов (например, как описано в [5]).

Во второй части строится матрица T порядка как транзитивное и рефлексивное замыкание отношения накрытия. Временная сложность этой части алгоритма равна $O(|V|^3)$.

Data: множество V=[0,1,…,n–1] значений текущего признака
Result: матрица B такая, что $B[j]$ - битовая строка для кодирования значения j
V = *topological_sort*(V); // топологическая сортировка
$\forall i \forall j[T[i][j] = false]$; // матрица порядка
for (*index*=0; *index*<*n*; ++*index*) **do**
  $T[V[index]][V[index]]$ = *true*;
  for (*indx*=0; *indx*<*index*; ++*indx*) **do**
    if ($V[indx] < V[index]$) **then**
      for (*ndx*=0; *ndx*<*n*; ++*ndx*) **do**
        $T[V[index]][ndx]$ |= $T[V[indx]][ndx]$;
      end
    end
  end
end
$\forall i[Del[i] = false]$; // удаляемые столбцы
for (*index*=2; *index*<*n*; ++*index*) **do**
  for (*indx*=1; *indx*<*index*; ++*indx*) **do**
    for (*ndx*=0; *ndx*<*indx*; ++*ndx*) **do**
      if ($T[][V[index]] == T[][V[indx]] \& T[][V[ndx]]$) **then**
        $Del[V[index]]$ = *true*;
      end
    end
  end
end
for (*ndx*=*index*=0; *ndx*<*n*; ++*index*) **do**
  ndx = *index*;
  while ($Del[ndx]$ && *ndx*<*index*) **do**
    ++*ndx*;
  end
  if (*ndx* < *n*) **then**
    for (*indx*=0; *indx*<*n*; ++*indx*) **do**
      $B[index][indx] = T[ndx][indx]$;
    end
  end
end

Алгоритм 1: Кодирование битовыми строками

Главная часть – третья – обнаружение лишних столбцов. Временная сложность этой части равна $O\left(|V|^4\right)$. Если структура для хранения временной матрицы T представляет собой *std::list<boost::dynamic_bitset<>>* (список столбцов), то операция в самом внутреннем цикле хорошо векторно распараллеливается современными компиляторами, чем сильно уменьшается реальное время работы.

Наконец, последняя часть алгоритма составляет кодировочную матрицу B из оставшихся столбцов (=бинарных признаков). Ясно, что временная сложность этой части равна $O\left(|V|^2\right)$.

Доказательство корректности этого алгоритма составляет утверждение следующей леммы:

Лемма 5. ***Для решетки ВКФ-кандидатов L, порождаемой формальным контекстом*** $\leq \subseteq L \times L$, ***образ признака*** $h(f) = \left\langle \{f\}', \{f\}'' \right\rangle$ ***является*** $\cup$***-разложимым элементом, если и только если найдутся такие два признака*** $f_1 < f$ ***и*** $f_2 < f$, ***что*** $\{f\}' = \{f_1\}' \cap \{f_2\}'$.

Доказательство. По Определению 4 для $\cup$-разложимого элемента $\left\langle \{f\}', \{f\}'' \right\rangle$ должна найтись такая пара $\langle A_1, B_1 \rangle$ и $\langle A_2, B_2 \rangle$, что $\left\langle \{f\}', \{f\}'' \right\rangle = \langle A_1, B_1 \rangle \cup \langle A_2, B_2 \rangle$.

Положим $f_1 = B_1$ и $f_2 = B_2$, тогда $\{f_1\}' = A_1$ и $\{f_2\}' = A_2$. По Теореме 1 и Лемме 1 имеем

$$\left\langle \{f\}', \{f\}'' \right\rangle = \left\langle A_1 \cap A_2, (B_1 \cup B_2)'' \right\rangle = \left\langle \{f_1\}' \cap \{f_2\}', (f_1 \cup f_2)'' \right\rangle .$$

Согласно Определениям 2 и 4 из Леммы 2 получаем $f_1 \in \{f_1\}'' \subset \{f\}''$, т.е. $f_1 < f$. Аналогично, $f_2 < f$.

Ранее в работе [6] идея этого алгоритма применялась для представления сложных структур значений признаков при описании медицинских данных битовыми строками и рассматривалось несколько типов структур, широко используемых в реальных экспериментальных исследованиях.

В настоящей работе представлен алгоритм, который расширяет этот подход к кодированию битовыми строками на самый общий случай (произвольной нижней полурешетки), а также доказана корректность предложенного алгоритма и получена оценка его вычислительной сложности.

* * *

Автор благодарит своих коллег по отделу 16 Отделения 1 ФИЦ ИУ РАН за поддержку и полезные дискуссии. Особая благодарность выражается студентам отделения интеллектуальных систем Российского государственного гуманитарного университета, которые выступали первыми слушателями и критиками описываемого подхода (в рамках курса «Теория сходства в интеллектуальных системах»).

СПИСОК ЛИТЕРАТУРЫ

1. Vinogradov D.V. VKF-method of hypotheses generation // Communications in Computer and Information Science. – 2014. – Vol. 436. – P. 237–248.
1. ДСМ-метод автоматического порождения гипотез: Логические и эпистемологические основания / ред.: В.К. Финн, О.М. Аншаков. – М.: URSS, 2009. – 432 с.
2. Ganter B., Wille R. Formal Concept Analysis / transl. from German. – Berlin: Springer–Verlag, 1999. – 284 p.
3. Davey B.A., Priestley H.A. Introduction to Lattices and Order. – 2-nd eds. – Cambridge: Cambridge University Press, 2002. – 298 p.
4. Кормен Т., Лейзерсон Ч., Ривест Р., Штайн К. Алгоритмы: построение и анализ. – 2-е изд. / пер. с англ. – М.: Вильямс, 2005. – 1296 с.
5. Панкратова Е.С., Виноградов Д.В. Формальное описание настройки интеллектуальных ДСМ-систем на область клинической и лабораторной диагностики // Научная и техническая информация. Сер. 2. – 2011. – № 9. – С. 1–5; Pankratova E.S., Vinogradov D.V. Formal Description of Adaptation of Intelligent JSM-Systems for Clinical and Laboratory Data Analysis // Automatic Documentation and Mathematical Linguistics. – 2011. – Vol. 45, № 5. – P. 213-217.

Материал поступил в редакцию 12.01.2018

Сведения об авторе

ВИНОГРАДОВ Дмитрий Вячеславович – кандидат физико-математических наук, старший научный сотрудник Федерального исследовательского центра «Информатика и управление» РАН и Российского государственного гуманитарного университета, Москва e-mail: vinogradov.d.w@gmail.com

УДК 004.65:004.774:004.82

Е.М. Бениаминов

Библиотеки онтологий в Веб: состояние и перспективы

Дается обзор использования компьютерных онтологий для связывания данных в Интернете и представления знаний. Рассматривается тенденция организации библиотек и серверов онтологий для совместной коллективной разработки онтологий и их использования. Библиотеки онтологий рассматриваются как открытый ресурс в Веб. Особое внимание уделяется проблемам формирования библиотек онтологий, их отладки и направлениям развития в этой области.

Ключевые слова: *онтологии, библиотека онтологий, представление знаний, связанные данные, открытый язык*

ВВЕДЕНИЕ

Термин «онтология» в информационных технологиях стал очень популярным. Почти на всех последних конференциях в области IT был раздел, посвященный онтологиям. Компьютерные технологии, в которых используются онтологии, становятся ключевыми в развитии современного Веба, аналитики и управления большими данными, машинного обучения, в приложениях в области медицины, науки, образовании, торговли, производстве, в музейном деле и т.д.

Последний наш обзор [1] по технологиям с использованием компьютерных онтологий был написан 10 лет назад, в нем давались следующие определения.

- Онтологии представляют собой спецификации на формальном языке, в которых фиксируются договоренности группы специалистов о том, что как называется в их области, какими свойствами обладают и каким соотношениям удовлетворяют.
- На логическом уровне каждой онтологии соответствует некоторая теория (сигнатура+аксиомы), а иногда и некоторая фиксированная модель (множества+операции+отношения). Вопросы к онтологии интерпретируются как запросы к соответствующей ей теории (модели).
- Онтологии, как правило, строятся по модульному принципу: при определении новой онтологии могут использоваться уже построенные ранее.
- Онтологии должны быть удобны для понимания специалистами и интерпретироваться компьютерными системами при использовании.

В настоящей статье мы продолжаем придерживаться этих определений. Обзор [1] заканчивался призывом к созданию открытых библиотек онтологий в Веб с открытыми языками формирования онтологий и запросов к ним.

За последние 10 лет появилось много новых библиотек и серверов онтологий. Некоторые из них специализированы и разрабатываются целенаправленно в ведомствах, фирмах и объединениях. Некоторые онтологии становятся ведомственными стандартами, например в медицине [2, 3], в музейном деле [4], в госучреждениях [5].

Уже в 2012 г. появился хороший обзор [6], название которого можно перевести как «Где публиковать и находить онтологии? Обзор онтологических библиотек». В нем приводится сравнение нескольких библиотек онтологий и показывается, что многие из них разработаны с разными целями использования, а также приводятся некоторые общие принятые требования к библиотекам онтологий в Веб.

ОНТОЛОГИИ ДЛЯ ОРГАНИЗАЦИИ СВЯЗАННЫХ ДАННЫХ В ВЕБ

Особое место занимают библиотеки онтологий, используемые в проекте открытых связанных данных, который был объявлен одним из создателей Веба Тимом Бернерсом-Ли в 2001 г. В эмоциональной видео-лекции «От гипертекстовой организации страниц и серверов открытых энциклопедий к открытым и связанным данным в Веб» [7], в которой говорится о развитии Веба в направлении открытых связанных данных, Тим Бернерс-Ли призывает участников Веба открывать свои данные (базы данных) для общего пользования, обеспечивая доступ к данным через онтологии, описывающие схемы их базы данных. При этом он на примерах демонстрирует общую выгоду от использования открытых данных через связывание данных в общую систему. Естественно, что онтологиям и библиотекам онтологий в этой системе отводится особая роль по обеспечению согласования данных из разных источников и по добыче знаний из больших объемов данных [8].

DBpedia [9] – это проект, направленный на извлечение структурированной информации из данных, созданных в рамках проекта Википедиа, и публикации её в виде доступном под свободной лицензией наборов данных. Проект был отмечен Тимом Бернерсом-Ли, как один из наиболее известных примеров реализации концепции связанных данных. Как отмечается в [9], по состоянию

на ноябрь 2016 г. базы данных DBpedia описывали уже более 6,6 млн сущностей, из которых 4,9 млн имеют аннотации, 1,9 млн имеют географические координаты. В целом, 5,5 млн источников Интернета расклассифицированы в соответствии с онтологией DBpedia, состоящей, в том числе, из 1,5 млн персоналий, 440 тыс. географических объектов, 139 тыс. музыкальных альбомов, 111 тыс. фильмов, 21 тыс. видеоигр, 286 тыс. организаций.

Как отмечено разработчиками DBpedia в [10], онтология DBpedia была создана вручную на основе наиболее часто используемых информационных подразделов в Википедии. Онтология в настоящее время охватывает 685 классов, которые образуют иерархию и описываются 2795 различными свойствами.

С выпуском DBpedia 3.5 была открыта библиотека Wiki [11], с помощью которой пользователи могут самостоятельно редактировать онтологии DBpedia и сопоставлять ее элементы с элементами Википедии приблизительно так же, как производится редактирование страниц в Википедии.

Выпуск DBpedia 2016-10 состоит из 13 млрд (2016-04: 11,5 млрд) информации в виде троек RDF, из которых 1,7 млрд (2016-04: 1,6 млрд) были извлечены из английского издания Wikipedia, 6,6 млрд (2016-04: 6 млрд) были извлечены из других изданий на разных языках изданий и 4,8 млрд (2016-04: 4 млрд) из Wikipedia Commons и Wikidata.

ТИПЫ И ЗАДАЧИ БИБЛИОТЕК ОНТОЛОГИЙ

В настоящее время библиотеки онтологий созданы для разных целей и, таким образом, обеспечивают разные функциональные возможности, имеют очень разный объем и могут быть специализированы по предметным областям. Поэтому пользователь, который хочет повторно использовать онтологию из множества сотен или тысяч онтологий, доступных в Интернете, должен не только использовать библиотеку онтологий, но и иметь средства поиска и анализа библиотек, чтобы разобраться в этом разнообразии и выбрать, какую онтологическую библиотеку использовать.

Как отмечено в [6], существуют онтологии, которые используются для описания данных в социальных сетях; есть онтологии, которые становятся стандартом для описания продуктов и услуг коммерческих организаций. Некоторые сообщества специалистов в конкретных областях деятельности достигли соглашения о применяемых у них онтологиях, обеспечив высокий уровень повторного использования этих конкретных онтологий. Например, многие биомедицинские исследователи используют генную онтологию для аннотирования своих данных. Точно так же сложился стандарт онтологии в области музейного дела. В таких областях создаются отдельные серверы и программные средства для коллективного формирования и поддержки стандарта онтологий в своих областях.

Есть более закрытые для редактирования библиотеки онтологий, например, библиотека онтологий SWEET [12] для наук об окружающей среде, разработанная в одной из лабораторий NASA. Онтологии SWEET написаны на языке OWL и являются общедоступными. SWEET-2.3 имеет модульную организацию и содержит около 6000 концепций в 200 отдельных онтологиях. Большинство пользователей использует эти онтологии как онтологии среднего уровня, и пополняет их онтологиями прикладной области для удовлетворения потребностей конечных пользователей.

Особый интерес представляет программная система OntoWiki [13], которая свободно распространяется и служит средством создания библиотек онтологий и связанных данных в стиле семантической Wiki. Это веб-приложение, написанное на PHP и использующее базу данных MySQL для коллективного использования. Система служит редактором онтологий на основе форм взаимодействия с пользователями. Имеется форма для формирования запросов на языке SPARQL к построенным и внешним онтологиям и модуль построения ответов на запросы.

Рассмотренные примеры библиотек онтологий показывают следующие проблемы, которые приходится решать при создании библиотек:

- большие онтологии и большие библиотеки онтологий;
- формирование сложных систем онтологий требует соответствующих средств опробования и отладки онтологий;
- для сложных онтологий полностью отделить непроцедурные и процедурные знания не удается (эффективность использования онтологий, прагматика);
- поддержка модульности построения онтологий и использования библиотек онтологий при создании новых онтологий;
- учет контекстности онтологий в библиотеке и взаимной противоречивости онтологий в различных контекстах;
- проблема интеграции онтологий, представленных на разных языках в разных логиках и моделях.

Для изучения последней проблемы некоторая инициативная группа создала из разнородных онтологий и отображений между ними пример библиотеки Open Ontology Repository [14]. Для согласования таких онтологий группа предлагает использовать современные алгебраические подходы к онтологиям.

БИБЛИОТЕКИ ОНТОЛОГИЙ С ИСПОЛЬЗОВАНИЕМ ФРАГМЕНТОВ ЕСТЕСТВЕННЫХ ЯЗЫКОВ ФОРМИРОВАНИЯ ОНТОЛОГИЙ И ЗАПРОСОВ К НИМ

Одна из важных проблем взаимодействия с библиотеками онтологий – это представление онтологий пользователям в удобном и понятном виде. Другая важная проблема – это программная поддержка процесса формирования новых онтологий специалистами в различных областях знаний, обеспечивающая использование уже существующих в библиотеке онтологий и процесс отладки проектируемой онтологии.

Решение этих проблем имеет несколько направлений. Первое – это использование графических редакторов для представления и формирование онтологий. На этом направлении все ещё лидирует по числу использований система Protégé [15], развивающаяся в соответствии с пожеланиями объединения пользователей Protégé. В настоящее время Protégé поставляется и в виде веб-сервера. При этом есть сервер, на котором хранятся онтологии, разработанные пользователями Protégé в виде библиотеки OWL-файлов [16]. От-

ладка сформированных онтологий производится с помощью программ логического вывода (резонеров). Они выделяют классы, которые не могут иметь экземпляры (пустые классы). С помощью резонеров строятся также ответы на запросы к онтологиям. Запросы могут быть написаны, например, на языке SPARQL.

Второе направление – это использование в библиотеке онтологий-шаблонов для формирования новых онтологий. К таким системам относится библиотека онтологий ODP [17]. Онтологии-шаблоны представляют собой фрагменты онтологий, хранящиеся в библиотеке и используемые в других онтологиях в виде модулей. В этом подходе предполагается построение больших онтологий из большого числа отлаженных онтологий-шаблонов и их согласований в рамках новой. Отладка онтологий производится по результатам тестирования по SPARQL запросам к проектируемой онтологии.

Третье направление – это использование ограниченного формализованного (контролируемого) фрагмента естественного английского языка для представления онтологий. К таким системам относится, например [18], в состав которой входит редактор онтологий Fluent Editor [19], позволяющий описывать онтологии на ограниченном фрагменте естественного английского языка. При этом онтология отображается графически и на языке OWL с использованием языка Semantic Web Rule Language (SWRL). В системе также есть Reasoner, с помощью которого проверяется правильность построения онтологии и строятся ответы на вопросы.

Онтологии, написанные на таком естественном языке, все равно читаются и понимаются с большим трудом. Писать на естественном языке специализированные онтологии также неестественно, как писать математические тексты на естественном языке без математических обозначений.

Следует заметить, что все эти направления не являются абсолютно новыми. Все это уже было (за исключением связанных данных) в системе Ontolingua, созданной и открытой для работы в Веб лабораторией KSL Стэнфордского университета в 1995 г. (система не поддерживается с 2010 г.). Более того, в системе Ontolingua была разработана громадная библиотека онтологий по многим областям знания и получена возможность создания на этой основе онтологий задач с помощью компьютерного моделирования.

ПЕРСПЕКТИВЫ БИБЛИОТЕК С ОТКРЫТЫМ ЯЗЫКОМ ФОРМИРОВАНИЯ ОНТОЛОГИЙ И ЗАПРОСОВ

Опыт проектирования онтологий показывает, что на пользовательском уровне применение графических методов представления онтологий и использование формализованного фрагмента естественного языка явно недостаточно. Почти в каждой области знаний при работе с онтологиями и их проектировании у пользователей возникает потребность в разработке и использовании специализированных для предметной области формализованных языков. Это ярко проявляется, когда представляются онтологии из математических областей знаний, но и в других областях знаний без использования языка предметной области текст онтологий становится трудночитаемым для специалистов. В предметных областях специализированный язык является более важным, чем естественный. А так как науки развиваются, развиваются языки, то и в системах представления онтологий должны быть средства ввода и использования новых языковых конструкций. Таким образом, хотелось бы, чтобы у пользователей для представления онтологий была возможность воспользоваться средствами открытого языка, и для формирования такого языка была соответствующая компьютерная поддержка.

Итак, будем исходить из следующих положений при развитии систем проектирования библиотек онтологий:

- так как онтология есть фиксация в формальном виде договоренностей группы специалистов определенной области о системе используемых ими понятий, их свойствах и аксиомах, то каждая система онтологий имеет смысл только для группы людей, принимающих эти договоренности (социальный характер онтологий);
- так как в онтологиях фиксируются договоренности специалистов, представлять онтологии должны специалисты предметных областей, поэтому язык представления онтологий должен быть удобен для этих специалистов;
- в каждой области знания при формировании понятий этой области формируются специализированные языки для работы с этими понятиями. Поэтому язык представления онтологий должен быть открытым для пользователей с возможностью его настройки для данной предметной области. При этом внутреннее представление онтологий должно быть стандартизованным для компьютерного использования и межмашинного обмена;
- так как науки и определенные представления в областях знаний меняются, то в компьютерных системах онтологий требуются средства поддержки целостности данных библиотеки онтологий при изменениях в языке и постепенном накоплении онтологий.

В связи с этим, предлагаются три принципа построения библиотек онтологий нового типа.

1. Онтологии строятся в стиле Wikipedia с поддержкой модульности, коллективной работы, версий и системы согласований (лучшие образцы WebProtégé и OntoWiki).
2. В системе поддерживается среда открытого языка работы с онтологиями, который, по мере пополнения базы онтологий, формируется самими пользователями.
3. Вместе с текстом онтологии в системе формируется внутреннее представление онтологии, которое используется при семантическом анализе выражений языка, при формировании ответов на запросы к онтологии и её отладке. При межмашинном обмене онтологиями и при использовании онтологий в приложениях возможен обмен в некотором стандарте, например в OWL.

В Российском государственном гуманитарном университете на кафедре математики, логики и интеллектуальных систем разработан прототип системы для формирования библиотек онтологий с откры-

тым языком представления онтологий [20]. Ведется сервер проекта [21]. Программы проекта, разработанные на основе программных средств Drupal и Visual Prolog 5.2, доступны на GitHub [22] под открытой лицензией GNU. Имеется руководство пользователя системой ЭЗОП [23].

ЗАКЛЮЧЕНИЕ

Представленный в статье обзор библиотек онтологий в Веб, позволяет утверждать, что при построении библиотек все ещё актуальными являются следующие задачи:

- использование Web 2.0-технологии для создания социальных сетей и сред в Web, для формирования, наполнения и использования самими пользователями библиотек онтологий;
- открытый язык представления онтологий для пользователя и стандартный – для внутреннего представления;
- предоставление пользователям Веб удобных средств модульного (с использованием чужих модулей) формирования внутреннего (семантического) представления данных своих страниц, онтологий схем своих баз данных и языка запросов к страницам и данным.

Полезно также использование алгебраического подхода к моделированию онтологий, как к средству для интеграции разнородных онтологий.

СПИСОК ЛИТЕРАТУРЫ

1. Бениаминов Е.М. Некоторые проблемы широкого внедрения онтологий в IT и направления их решений // Труды Симпозиума «Онтологическое моделирование». – М.: ИПИ РАН, 2008. – С.71-82. – URL: http://beniaminov.rsuh.ru/BeniaminovOntoNew.pdf (дата обращения: 31.12.2018).
2. Открытые биомедицинские онтологии – URL: https://ru.wikipedia.org/wiki/Открытые биомедицинские онтологии (дата обращения: 31.12.2018).
3. The National Center for Biomedical Ontology. – URL: https://www.bioontology.org/ (дата обращения: 31.12.2018); The OBO Foundry. – URL: http://obofoundry.org/ (дата обращения: 31.12.2018).
4. The CIDOC Conceptual Reference Model (CRM). – URL: http://www.cidoc-crm.org/ (дата обращения: 31.12.2018).
5. Ontologies for e-Government. – URL: http://www.oegov.us/ (дата обращения: 31.12.2018).
6. d'Aquin M., Noy N.F. Where to Publish and Find Ontologies? A Survey of Ontology Libraries // Web Semantics: Science, Services and Agents on the World Wide Web. – 2012. – Vol. 11. – P. 96-111. – URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3293483/ (дата обращения: 31.12.2018).
7. Berners-Lee T. The next web // TED2009. – URL: https://www.ted.com/talks/tim_berners_lee_on_the_next_web (дата обращения: 31.12.2018).
8. Abele A., McCrae J.P., Buitelaar P., Jentzsch A., Cyganiak R. Linking Open Data cloud diagram 2017. – URL: http://lod-cloud.net (дата обращения: 31.12.2018); Linked Data. – URL: https://en.wikipedia.org/wiki/Linked_data; http://linkeddata.org/ (дата обращения: 31.12.2018).
9. DBpedia. – URL: http://wiki.dbpedia.org/; http://wiki.dbpedia.org/datasets/dbpedia-version-2016-10 (дата обращения: 31.12.2018).
10. Ontology of DBpedia. – URL: http://wiki.dbpedia.org/services-resources/ontology (дата обращения: 31.12.2018).
11. DBpedia Mappings Wiki. – URL: http://mappings.dbpedia.org/index.php/Main_Page (дата обращения: 31.12.2018).
12. Semantic Web for Earth and Environmental Terminology (SWEET). – URL: https://sweet.jpl.nasa.gov (дата обращения: 31.12.2018).
13. Semantic Data Wiki and Linked Data Publishing Engine. – URL: https://ontowiki.net/; https://docs.ontowiki.net/ (дата обращения: 31.12.2018).
14. Open Ontology Repository (OOR) Initiative - Home Page. – URL: http://www.oor.net/; http://ontologforum.org/index.php/OpenOntologyRepository; http://ontolog.cim3.net/wiki/OpenOntologyRepository.html#nid17YN (дата обращения: 31.12.2018).
15. A free, open-source ontology editor and framework for building intelligent systems Protégé. – URL: https://protege.stanford.edu/ (дата обращения: 02.02.2018).
16. Protege_Ontology_Library. – URL: https://protegewiki.stanford.edu/wiki/Protege_Ontology_Library (дата обращения: 02.02.2018).
17. Ontology Design Patterns. org (ODP). – URL: http://ontologydesignpatterns.org (дата обращения: 02.02.2018).
18. Cognitum. – URL: http://www.cognitum.eu/ (дата обращения: 02.02.2018).
19. Fluent Editor for PC. – URL: http://www.cognitum.eu/semantics/FluentEditor/ (дата обращения: 02.02.2018).
20. Web-сервер онтологий системы ЭЗОП. – URL: http://ontoserver.rsuh.ru/ezop/ (дата обращения: 02.02.2018).
21. Сервер проекта системы ЭЗОП. – URL: http://ezop-project.ru/ (дата обращения: 02.02.2018).
22. Тексты программ системы ЭЗОП. – URL: https://github.com/beniaminov/ezop_server; https://github.com/beniaminov/WebEzop (дата обращения: 02.02.2018).
23. Бениаминов Е.М. Работа в системе коллективного формирования библиотек онтологий ЭЗОП (руководство пользователя). – М.: РГГУ (препринт), 2015. – 49 с. – URL: http://beniaminov.rsuh.ru/User_guideEzop.pdf (дата обращения: 02.02.2018).

Материал поступил в редакцию 05.02.18.

Сведения об авторе

БЕНИАМИНОВ Евгений Михайлович – доктор физико-математических наук, профессор, заведующий кафедрой математики, логики и интеллектуальных систем Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Российский государственный гуманитарный университет» (РГГУ), Москва
e-mail: ebeniamin@yandex.ru

УДК 004.65.056

В.В. Грибова, А.В. Иванова

Автоматизация выбора средств защиты информационных систем на основе онтологического подхода*

Рассмотрены этапы построения системы защиты информации, проанализированы достоинства и недостатки существующих методов их реализации. Предложена концепция инструментального комплекса выбора средств защиты, для осуществления которой предлагается использовать облачные технологии и онтологический подход к формированию информационных ресурсов, что обеспечит гибкость и оперативную расширяемость системы без необходимости изменения программного кода.

Ключевые слова: *базы данных, базы знаний, онтологии, информационные системы, информационная безопасность, система защиты*

ВВЕДЕНИЕ

При организации информационной безопасности систем выбор составляющих ее компонентов зависит от результатов выполнения предшествующих этапов: определения характеристик защищаемой системы, выделения актуальных угроз безопасности и соответствующих им требований безопасности. Требования безопасности, в свою очередь, могут характеризоваться различными критериями: необходимостью обеспечения конкретного уровня защищенности, нейтрализацией актуальных уязвимых мест, исполнением законодательных нормативов и т.д. На сегодняшний день требования безопасности в полной мере могут быть определены экспертным путем, а также частично при помощи аналитических комплексов [1].

Использование экспертного подхода предполагает привлечение специалистов для описания системы, определения её характеристик и классификации, выявления актуальных угроз безопасности, а также для последующего выбора средств защиты, их установки и настройки. Все этапы организации системы защиты выполняются в «ручном» режиме и предусматривают непрерывное сопровождение экспертом.

Привлекаемый специалист должен обладать навыками и знаниями, позволяющими ему не только осуществлять паспортизацию информационной системы, но и выстраивать систему защиты в соответствии с динамично развивающейся законодательной базой, а также подстраивать ее под ежедневные потребности предприятия. Это позволяет сочетать меры законодательного, административно-организационного, и программно-технического характера. Однако, как отмечено в [2–4], такой подход обладает рядом недостатков, основными из которых являются: отсутствие достаточного количества экспертов (сложность анализа и оценки информационной безопасности приводит к тому, что человеку необходимо тратить очень много времени для того, чтобы стать экспертом), высокая стоимость оказываемых услуг, а также низкая оценка адекватности экспертного анализа.

Применение аналитических комплексов для выявления уязвимости системы позволяет автоматизировать процесс определения требований безопасности. Однако, для выбора средств защиты, их установки и настройки, а также для последующего сопровождения системы защиты требуется привлечение экспертов. Данный подход имеет преимущественно программно-технический характер и не предусматривает поддержку и выполнение законодательных, организационно-правовых, административных мер в полном объеме. При нем лишь определяются уязвимые места исследуемых систем, в отдельных случаях выдаются рекомендации по их нейтрализации [1, 5, 6].

Комплексное решение проблемы обеспечения информационной безопасности, рациональное сочетании законодательных, административно-организационных и программно-технических мер, обязательное следование промышленным, национальным и международным стандартам – это тот фундамент, на котором строится вся система защиты. При выборе способа

* Работа выполнена при частичной финансовой поддержке РФФИ (гранты 16-07-00340, 18-07-01079) и программы "Дальний Восток" (проект 18-5-078).

объединения средств защиты в единую систему также могут применяться два метода [7]: «фрагментарный», основанный на использовании набора разнородных средств защиты информации (СЗИ) [4] или «комплексный» - использующий единое многофункциональное решение одного производителя [7, 8]. Наиболее распространен первый подход, предусматривающий установку и настройку комплекса разнородных средств защиты. Однако, в связи с его недостатками [9] для организации такого комплекса необходимо привлекать высококвалифицированных специалистов, как для создания, так и для дальнейшего сопровождения системы [10].

Любая организация системы защиты, не предусматривающая её комплексность, является нерациональной ввиду недостаточности принимаемых мер. Однако, как показал анализ, комплексный автоматизированный подход для выполнения вышеуказанных задач отсутствует, имеются решения, реализующие автоматизированное выполнение отдельных этапов.

Цель работы – описание программно-информационного комплекса для автоматизированного выбора СЗИ (с учетом законодательных, административно-организационных и программно-технических мер, промышленных, национальных и международных стандартов), удовлетворяющего основным принципам организации системы защиты информации. Объектом исследования являются информационные системы, а также методы и средства обеспечения их защиты в соответствии с требованиями безопасности информации.

ОБЗОР РЕШЕНИЙ, РЕАЛИЗУЮЩИХ В АВТОМАТИЗИРОВАННОМ РЕЖИМЕ ВЫПОЛНЕНИЕ ОТДЕЛЬНЫХ ЭТАПОВ ОРГАНИЗАЦИИ СИСТЕМЫ ЗАЩИТЫ

Первым этапом при выборе средств защиты является паспортизация и классификация системы, определение требований безопасности на основе различных критериев, в том числе актуальных угроз безопасности, для конкретной системы. В общем случае выполнение данной задачи основывается на использовании справочной и методологической литературы [11-13] и ее последующем анализе с целью определения требований к системе. Для автоматизации этого этапа некоторые авторы предлагают применять методы искусственного интеллекта. Так, Т.Н. Ворожцова, в своих работах [14,15] рассматривает построение иерархии системы фундаментальных понятий в области кибербезопасности на основе онтологий. В качестве источников терминологии и взаимосвязей между понятиями для наполнения баз данных взяты нормативно-методические документы. Также в качестве основы при управлении рисками для объединения в систему защиты программных и аппаратных комплексов согласно международному стандарту 27001 зарубежными авторами используются онтологии [16, 17]. Рассматриваемые авторами комплексы поддерживают протокол автоматизации управления данными безопасности *SCAP* в рамках программы автоматизации информационной безопасности *ISAP* [18], а потому речь идет об их единой онтологии при организации данных безопасности.

Развитие интеллектуальных технологий в области кибербезопасности осуществляется за счет построения экспертных комплексов, включающих в себя базы данных, базы знаний в форме правил (которые на основе опроса пользователя выявляют и анализируют уязвимые места описанной системы [19, 20]), использования экспертных систем [21, 22], а также использования онтологий для организации экспертных знаний [23].

Однако существующие решения для паспортизации, классификации и определения требований безопасности обладают рядом недостатков и во многом отступают от базовых фундаментальных принципов организации программных комплексов управления информационной безопасностью [9]. Анализ показал, что основные проблемы заключаются в следующем:

- применяется не общепринятая терминология;
- представленные онтологии отражают лишь один из начальных этапов решения задачи классификации, так как затрагивают только общие аспекты безопасности;
- информационные ресурсы не обладают универсальностью по отношению к разным типам информационных систем;
- для масштабирования баз данных и баз знаний, а также для изменения понятийного аппарата необходимо изменять программный код;
- системы опроса пользователя работают только с логическими вопросами, что не позволяет детально описать количественные, диапазонные, изменяемые характеристики системы;
- требуется наполнение базы данных бесконечно большим количеством логических вопросов для создания точного описания;
- для определения требований безопасности не существует решений на основе нормативно-методической базы.

Второй этап включает в себя определение уязвимых мест, угроз, средств и методов защиты для конкретной информационной системы. Как показал обзор литературы, существуют автоматизированные решения, позволяющие определить достаточный набор средств защиты для обеспечения безопасности конкретной информационной системы [24, 25]. Входными параметрами для таких систем служат результаты анализа угроз и оценки уровня защищенности. Для реализации этого этапа в своих работах Е.А. Рахимов [26], И.В. Машкина [24] используют механизм нечеткого логического вывода на основе данных о технических характеристиках средств защиты, декларируемых разработчиками. Однако, несмотря на многолетние исследования государственных институтов в области защиты информации, авторы делают выводы, что стандарты информационной безопасности и государственные методические документы не формируют конкретных подходов к управлению безопасностью. Считается, что они определяют лишь функциональные требования в отношении средств защиты и не предлагают методик сравнительного анализа различных комплексов средств защиты в целях выявления наиболее рационального варианта. По этой причине авторы не учитывают их при определении требований к системе.

Для выбора средств защиты используется «Трехрубежная модель защиты»: периметр, сегмент, компьютер. В первую очередь, в зависимости от функционала, разрабатывается множество вариантов наборов СЗИ, которые задаются морфологической матрицей. Затем наборы уточняются для каждого из рубежей. Заполняются вспомогательные матрицы, в которых отмечаются совместимые друг с другом программно-аппаратные средства. На завершающем этапе генерируется множество решений по выбору вариантов набора средств защиты, осуществляется усечение этого множества до подмножества вариантов набора из совместимых между собой программно-аппаратных продуктов.

В системе интеллектуальной поддержки, разработанной под руководством Машкиной И.В. [24], *рациональные* решения выбираются на основе использования многокритериального сравнительного экспертного анализа, в результате которого в заданном экспертом множестве выявляются подмножества наилучших по критериям предпочтения вариантов наборов, из которых формируется рациональный комплекс средств защиты. Критерии качества, определяющие рациональность средств защиты, по иерархии «защищенность» делятся на две группы: критерии обеспечения эффективности оперативных методов защиты и критерии функциональной пригодности. Критерии качества по иерархии «издержки» делятся также на две группы: в первую включена стоимость соответствующего средства защиты, число пользователей по одной лицензии и другие возможные экономические издержки; ко второй группе издержек относятся функциональные издержки, такие, например, как падение производительности информационной системы при использовании данного СЗИ. Для приобретения знаний в рамках разработанного программного комплекса организуется взаимодействие эксперта с автоматизированной системой, в процессе которого эксперт заполняет предложенные ему разработанные поля знаний.

Подобные программные средства позволяют осуществлять анализ уязвимых мест и выбор методов защиты. Однако они обладают *рядом недостатков*:

- в случае изменения технических характеристик СЗИ (новая версия, стоимость, функционал и т.д.) эксперту необходимо осуществить переоценку матриц наборов СЗИ, в том числе для каждого из рубежей, а также матриц совместимости СЗИ и снова осуществить многокритериальный сравнительный экспертный анализ наборов СЗИ;
- не принимаются во внимание законодательные требования к безопасности информации, которые формализуют критерии, учитывая не только тип объекта информатизации (компьютер, сегмент сети, сеть), но и среду передачи данных, количество пользователей в системе, типы идентификаторов и другие характеризующие признаки;
- как отмечает автор, на момент исследования неясность способа определения вероятности угроз и уязвимых мест являлась основной проблемой при получении количественной оценки риска нарушения информационной безопасности. Методика для этого существовала, но была предназначена только для служебного пользования. Пометка «для служебного пользования» снята Решением ФСТЭК России от 16 ноября 2009 г. Однако на текущий момент «Методика определения актуальных угроз безопасности персональных данных при их обработке в информационных системах персональных данных» [13], уже устарела, а новый проект находится в разработке ФСТЭК;
- уменьшение количества экспертов, имеющих возможность наполнять базы знаний, прямо пропорционально увеличивает объем знаний необходимых одному эксперту. В условиях динамично меняющихся обстоятельств, влияющих на безопасность (законы, СЗИ, угрозы, уязвимости и т.д.) следует предоставить возможность узконаправленным экспертам вносить в систему знания в касающейся их части.

Третий этап – объединение модульных средств в единую систему защиты информации.

В области построения и управления комплексными СЗИ можно выделить фундаментальные и узконаправленные исследования. Фундаментальные исследования [3, 4, 26] посвящены общим принципам децентрализованной системы защиты, ее достоинствам и недостаткам. Отмечается, что комплексное использование предполагает согласование разнородных средств при построении целостной системы защиты. При этом стандартный набор модульных средств содержит следующие компоненты: средства обеспечения надежного хранения информации, средства защиты информации от несанкционированного доступа (СЗИ от НСД), межсетевые экраны, антивирусные средства, криптошлюзы, средства защиты от *DDoS*, средства централизованного управления системой и др.

К узконаправленным исследованиям относятся работы, описывающие конкретные решения и возможности объединения в единую комплексную систему разнородных средств защиты, производимых различными разработчиками. Здесь необходимо отметить труды Сухарева С.В., Лыдина С.С., Макейчика Ю.С. [27], Рахимова Е.А. [25], Машкиной И.В. [24]. Коммерческие компании [28] также изучают вопросы централизованного управления однородными (несколько межсетевых экранов) и разнородными (комплексное) СЗИ. Последние выпускаются с единой консолью управления своими программными или аппаратными модулями: СЗИ от НСД, межсетевым экраном, средством обнаружения вторжений и другими. Системы централизованного управления однородными средствами защиты последние несколько лет производятся в составе комплектов программного обеспечения (ПО) СЗИ и представляют собой отдельный модуль, который можно настроить при установке, указав параметры центра управления [8, 28, 29]. Некоторые средства защиты предусматривают многоуровневую централизованную систему. К примеру, в пределах предприятия для управления СЗИ от НСД используется «домен безопасности», а для централизованного управления настройками на региональном уровне (для управления «доменами безопасности») используется «лес» безопасности [30].

При решении задачи объединения разнородных децентрализованных средств может возникнуть проблема, связанная с объединением разрозненной ин-

формации. Ее можно решить путем сопоставления и совместного использования данных безопасности из различных средств защиты, используемых в информационной системе. Для этого требуется система управления компонентами мониторинга и управления. В [9] отмечается, что с этой целью могут быть использованы внешние системы управления.

Один из вариантов, позволяющих централизованно выполнять оперативный мониторинг (например, видеть, какие серверы в данный момент работают), мониторинг производительности, управление рабочими станциями (инвентаризация, управление ПО, управление идентификационными данными пользователей, управление хранением данных, и т.д.) – *IBM Tivoli* [31]. При совместном применении *Tivoli* и системы централизованного управления и мониторинга СЗИ от НСД последняя используется в качестве единого консолидированного источника данных о событиях информационной безопасности, зафиксированных СЗИ от несанкционированного доступа.

В работах [24, 25, 32, 33] описаны другие примеры внедрения внешней управляющей системы. В ней используется не только собственный механизм анализа событий и генерации данных аудита, но и внутренний механизм формирования управляющих команд для средств защиты.

Следует отметить, что подход, основанный на интеграции дополнительной системы управления *Tivoli* в штатную ОС, для управления настройками безопасности СЗИ *не рационален* по следующим причинам:

- внутренняя система аудита дублирует штатный механизм сбора событий, с которым, как правило, совместимы функции экспорта и импорта данных большинства систем защиты информации;
- высокая стоимость промышленных решений;
- в большинстве случаев отсутствие прямой совместимости со средствами защиты информации в части импорта, экспорта данных о событиях безопасности;
- повышение нагрузки на вычислительные мощности информационной системы;
- необходимость установки программного комплекса на платформу;
- отсутствие возможности масштабирования системы пользователем в режиме реального времени.

Исходя из вышеизложенного, следует отметить, что существует потребность в организации средств и методов создания такой системы защиты, при которой администратор информационной системы, обладая минимальными экспертными навыками в области защиты информации, был бы способен создать комплексную систему защиты, соответствующую конкретной информационной системе, пользовательским ожиданиям, а также нормативам в области безопасности информации для этой системы. Результат работы таких средств следует воспроизводить в понятной пользователю терминологии, исключая, тем самым, необходимость приглашения квалифицированных специалистов для его интерпретации. Комплексного решения, реализующего выполнение этих задач в автоматизированном режиме, на сегодняшний день не существует.

КОНЦЕПЦИЯ ВЫБОРА СРЕДСТВ ЗАЩИТЫ

Для выбора средств защиты предлагается комплексный автоматизированный подход, включающий в себя выполнение каждого из этапов организации системы защиты.

Каждый этап состоит из одной или нескольких подзадач, решение которых необходимо [34, 35] либо для перехода к следующему этапу, либо для решения конечной задачи. Так, на первом этапе создается паспорт информационной системы (ИС) и определяется ее класс; на втором этапе на основании класса ИС формируется набор требований безопасности; третий этап – требования безопасности позволяют определить функции средств защиты, набор программных средств в которых они реализованы, а также предоставить администратору набор рекомендаций по установке и настройке выбранного множества СЗИ. Каждый из этапов реализуется своими программными средствами, которые можно разделить на две группы: сервисы, предназначенные для администратора ИС и сервисы разработки и сопровождения, предназначенные для инженеров знаний и экспертов. На рис. 1 показаны сервисы для каждого этапа, а также наборы информационных ресурсов, которые они используют.

При реализации такой комплексной системы необходимо соблюдать следующие требования: гибкость и расширяемость системы без изменения ее программного кода (поскольку стандарты информационной безопасности постоянно изменяются), а также ее доступность, простота использования администраторами, что позволит не прибегать к услугам специалистов по компьютерной безопасности .

Исходя из этих требований, при формировании информационных ресурсов предлагается использовать онтологический подход. Это обеспечит гибкость и оперативную расширяемость системы. При этом изменять программный код не потребуется [36]. Также для обеспечения легкого, кроссплатформенного доступа как к пользовательским средствам, так и к средствам администрирования системы предлагается применять облачные технологии.

В рамках предлагаемой концепции информационные ресурсы представляют собой базы данных и базы знаний, а также их онтологии. Информационные ресурсы могут быть отредактированы или дополнены через соответствующие средства разработки и отладки. Связь онтологий информационных ресурсов между собой изображена на рис. 2.

Концепция механизма выбора средств защиты предусматривает формирование промежуточных результатов (рис. 3). Таким образом, пользователь, определив лишь характеристики своей информационной системы в предлагаемых ему формализованных (общепринятых, стандартизированных) терминах, получает информацию не только о средствах защиты, но и информацию о классификации его информационной системы, требованиях безопасности, необходимых функциях системы защиты для обеспечения безопасности, всех средствах защиты, включающих в себя эти функции, конечный набор средств защиты с конкретными рекомендациями по совместной настройке.

Аналогично способу взаимодействия пользователя с сервисом предлагается организовать взаимосвязь инженеров знаний для формирования и обновления онтологий и баз знаний, а также предметных специалистов для наполнения баз данных. Координацию работы рекомендуется осуществлять через средства разработки и сопровождения: редакторы онтологий, баз знаний и баз данных. В качестве предметных специалистов могут выступать как заинтересованные лица (разработчики и производители ПО), так и эксперты в предметных областях (рис.4). Эксперты совместно с инженерами знаний могут принимать участие в редактировании онтологий информационных ресурсов.

Этапы	Сервис администратора информационной системы	Средства разработки и сопровождения		Онтологии	Базы данных и базы знаний
Паспортизация и классификация	Создание паспорта информационной системы	Редактор характеристик информационной системы	Редактор онтологий	Онтология характеристик информационной системы	БД "Характеристики пользовательской системы"
	Сервис классификации информационной системы	Редактор классов		Онтология знаний о классификации	БД "Классы информационных систем"
Определение требований безопасности	Сервис определения требований безопасности	Редактор требований		Онтология требований	БД "Требования"
Выбор средств защиты информации	Сервис определения функций, реализующих требования безопасности	Редактор возможных функций средств защиты информации		Онтология возможных функций средств защиты информации	БД "БД Возможные функции средств защиты"
	Сервис определения средств защиты информации, содержащих функции	Редактор средств защиты информации		Онтология средств защиты информации	БД "Средства защиты информации"
	Сервис выдачи рекомендаций	Редактор рекомендаций		Онтология БЗ Рекомендации	БЗ "Рекомендации"

Рис. 1. Схема соответствия этапов выбора средств защиты информации, набора сервисов и информационных ресурсов

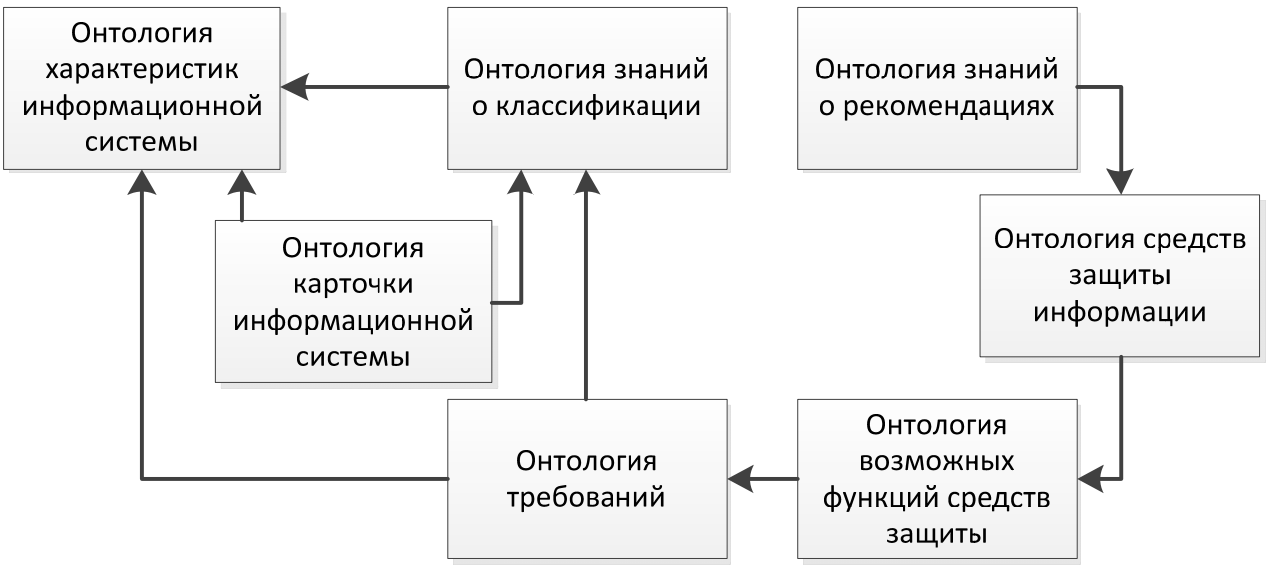

Рис. 2. Взаимодействие онтологий информационных ресурсов

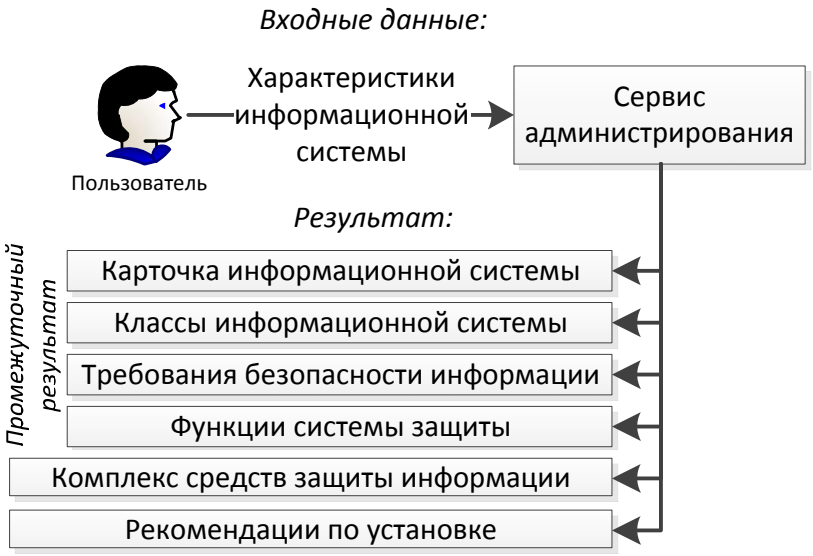

Рис. 3. Формирование результатов

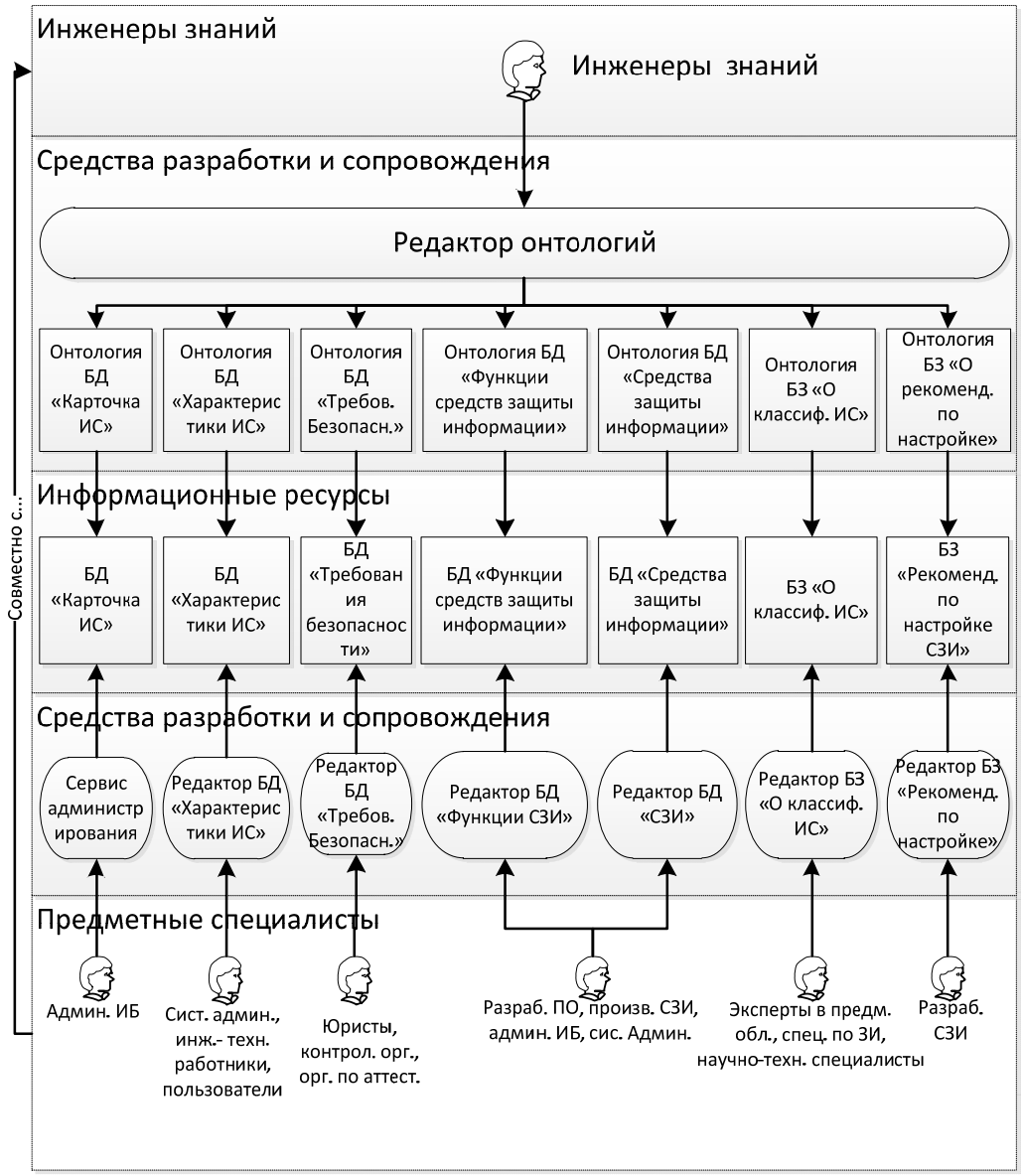

Рис. 4. Организация наполнения информационных ресурсов

Несмотря на имеющуюся возможность редактирования онтологий, предполагается, что они создаются однажды, являясь продуманными и универсальными. Редактирование может потребоваться, если онтология была спроектирована неправильно или система знаний изменилась настолько, что изменилась их структура. Для реализации программного комплекса выбрана платформа *IACPaaS* [37].

ИНФОРМАЦИОННЫЕ РЕСУРСЫ ПРОГРАММНОГО КОМПЛЕКСА ВЫБОРА СРЕДСТВ ЗАЩИТЫ

На этапе паспортизации и классификации первым функциональным элементом является задание пользователем характеристик информационной системы. Описательный этап, как правило, реализует эксперт.

Однако, как отмечалось ранее, такой подход обладает значительными недостатками. По этой причине предлагается автоматизировать процесс выбора средств защиты уже на этом этапе, а характеристики системы уточнять на основе выбора из БД «Характеристики» множества предлагаемых формализованных допустимых параметров.

В процессе работы взаимодействие с пользователем осуществляется через интерфейс платформы *IACPaaS*. Промежуточный результат формируется в виде «Паспортной карточки информационной системы», содержащей обобщенную информацию, представленную в терминологических понятиях. На основе формализованных характеристик определяется класс системы.

Согласно общему алгоритму работы модуля выбора средств защиты вырабатываются требования по защите информации. Каждая характеристика определяет одно или несколько требований к системе. Формально это можно представить следующим образом: «информационная система обрабатывает иные категории персональных данных сотрудников оператора или иные категории персональных данных менее чем 100000 субъектов персональных данных, не являющихся сотрудниками оператора». Для реализации на платформе условно это можно записать так: ((A∈1 и B∈2) и B∈3) или ((A∈1 и B∈3) и C<100000). После реализации оно будет прочитано следующим образом: ((«вид обрабатываемой информации» ∈ «персональные данные» И «тип персональных данных» ∈ «иные категории персональных данных») И «тип персональных данных» ∈ «персональные данные работников оператора») ИЛИ ((«вид обрабатываемой информации» ∈ «персональные данные» И «тип персональных данных» ∈ «Персональные данные субъектов, не являющихся работниками оператора») И «Количество субъектов ПДн» ∈ <100000).

Структура базы знаний и баз данных определяется онтологией. На основании требований безопасности информации к конкретной ИС определяются функции средств защиты информации, формируется набор функций, необходимых для конкретной ИС. Следует учесть, что, как и в случае с характеристиками, описанном выше, одна функция может реализовывать одновременно исполнение нескольких требований. Также возможно, что для исполнения одного требования необходимо активировать несколько функций.

Набор функций, необходимых к реализации, определяют средства защиты из БД «Средства защиты», в которых указанные функции реализованы. Помимо функций средств защиты можно внести в базу данных специфичные дополнительные характеристики, такие как стоимость или явный запрет (предпочтение) для конкретного средства защиты, для того, чтобы отображаемая информация по результатам работы модуля выбора средств защиты была более полезна и отвечала дополнительным критериям пользователя. Все функции описываются в соответствии с онтологией.

Результат работы модуля выбора средств защиты напрямую зависит от промежуточных результатов на этапе паспортизации (классификации). Предпочтительно использовать единый интерфейс для отображения обоих типов результатов: промежуточного и итогового. В качестве такового используется редактор платформы *IACPaaS*.

Таким образом, все информационные ресурсы, включающие в себя базы данных и базы знаний, объединены в единую систему (рис. 5), результатом функционирования которой является выбор средств защиты с учетом законодательных, административно-организационных и программно-технических мер защиты информации. Для минимизации затрачиваемых ресурсов при организации баз данных вместо клонирования (дублирования) идентичных блоков в каждой базе используются ссылки на существующие понятия между базами данных.

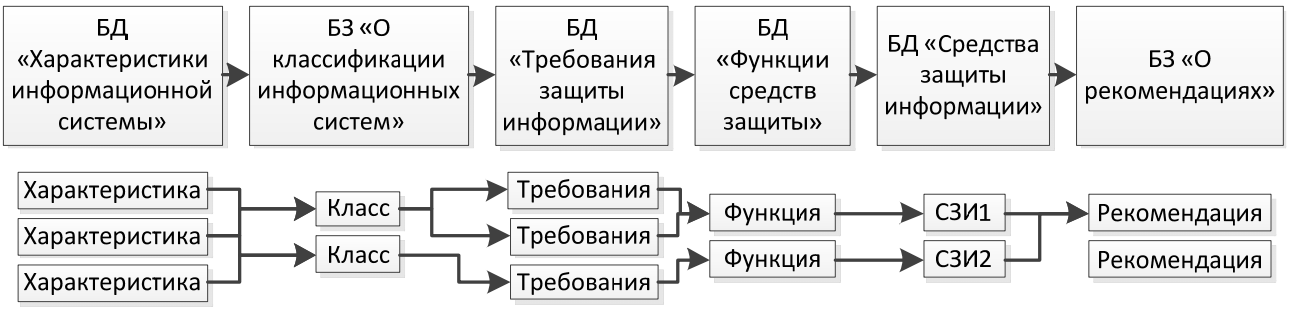

Рис. 5. Взаимосвязи между информационными ресурсами

Предложенный механизм организации системы защиты информации позволяет применять знания экспертов централизованно лишь для наполнения баз данных и баз знаний, а также для редактирования онтологий. Исходя из оптимизации использования трудовых ресурсов экспертов, их необходимо привлекать только для решения указанных задач, а сопоставление данных и знаний выполняет решатель. Это способствует созданию системы защиты без необходимости привлечения квалифицированных специалистов на каждом этапе, исключает связанные с этим недостатки комплексного подхода к построению систем защиты.

ОБНОВЛЕНИЯ И МАСШТАБИРОВАНИЕ

Поддержание актуальности «знаний модуля» предусмотрено за счет наполнения соответствующих баз данных и баз знаний серверной части инструментального комплекса [9]. Централизованное внесение изменений осуществляется экспертами, разработчиками ПО и другими заинтересованными специалистами. Такую процедуру предлагается осуществлять через Интерфейс редактора платформы *IACPaaS*. Это обеспечит возможность дистанционного своевременного наполнения системы актуальной нормативно-методической информацией, данными о новейших средствах защиты, предполагаемых пользовательских требованиях и т.д.

Концепция комплекса предусматривает интуитивно понятный для администратора алгоритм управления и сопровождения системы. Этот алгоритм через интерфейс редактора *IACPaaS* может осуществлять следующие функции:

- классификация системы (паспортизация);
- определение требований к системе;
- определение функционала системы защиты;
- определение конкретных средств защиты.

При необходимости поддерживается возможность внедрения графического интерфейса.

ЗАКЛЮЧЕНИЕ

На сегодняшний день развитие технологий систем принятия решений основано на организации и использовании таких общих банков данных, как: системы централизованного обновления антивирусных сигнатур [38], международные и российский банки угроз информационной безопасности [39, 40], облачные данные для эвристического анализа систем обнаружения вторжений [41] и др.

Централизованные банки данных позволяют принимать оперативные решения в самых различных областях на основе наиболее актуальной ключевой информации [42, 43]. Наполнение таких банков осуществляется экспертами с многолетним опытом и квалификацией, а в отдельных случаях в разработке банков данных участвуют международные эксперты. Централизованное управление ключевой информацией для систем принятия решений позволяет поддерживать ее высокую актуальность, оперативно исключать и исправлять ошибочные данные, поддерживать работоспособность множества клиентских платформ, используя один централизованный источник ключевой информации.

Анализ литературы показал, что в области информационной безопасности широко используются централизованные банки данных, однако, главной областью их применения является мониторинг и анализ угроз информационной безопасности. В области технологий организации средств защиты информации не предусмотрено ни централизованное использование банков информации, ни автоматизация указанного процесса – системы защиты разрабатываются при непосредственном физическом участии экспертов [44]. Предложенные в статье концепция и технология реализации механизма выбора средств защиты опираются на современные направления развития систем принятия решений в области информационной безопасности, такие как: централизованные банки информации, минимизация участия экспертов в разработке систем безопасности, онтологический подход к организации экспертных знаний, кроссплатформенность, облачная организация вычислительных ресурсов.

СПИСОК ЛИТЕРАТУРЫ

1. Сетевой сканер «Ревизор Сети 2.0» / Центр безопасности информации, 2016. – URL: http://www.cbi-info.ru/groups/page-356.htm
2. Созинова Е.Н. Применение экспертных систем для анализа и оценки информационной безопасности // Молодой ученый. – 2011. – Т.1, №10. – С. 64-66.
3. Семененко В.А. Информационная безопасность: учеб. пособие. – 4-е изд. – М.: МГИУ, 2010. – С. 75 – 78.
4. Шаньгин В.Ф. Защита информации в компьютерных системах и сетях. – М.: ДМК Пресс, 2012. – С. 243 – 246.
5. Сканер уязвимостей X-Spider / Positive Technologies, 2016. – URL: https://www.ptsecurity.com/ru-ru/products/xspider/
6. Системы обнаружения вторжений, сканеры безопасности / Securitylab.ru, 2016. – URL: http://www.securitylab.ru/software/1222/
7. «Обеспечение информационной безопасности сетей. Способы обеспечения информационной безопасности» / YourPrivateNetwork. Лаборатория сетевой безопасности, 2009. – URL: http://ypn.ru/146/securing-networking-information/
8. Код безопасности. Security Studio Endpoint Protection. Сертифицированная защита компьютера от сетевых вторжений, вредоносных программ и спама // Код безопасности, 2015. – URL: http://www.securitycode.ru/products/security_studio_endpoint_protection/
9. Грибова В.В., Иванова А.В. Концепция инструментального комплекса для построения системы защиты информационных систем // Открытые семантические технологии проектирования интеллектуальных систем. – 2016. – № 6. – С. 45 – 50.
10. Гайкович В., Ершов Д. Организационные меры обеспечения защиты информации. – URL:

http://itsecurity.ru/study-center/publication/organizatio-nal-measures-for-protection-of-the-information.php (дата обращения 18.05.2017)
11. Консультант Плюс – Надежная правовая поддержка, 2016. – URL: http://www.consultant.ru/
12. «Книги по информационной безопасности, криптографии хакингу» / Безопасник, 2016. – URL: http://bezopasnik.org/article/book/index.htm
13. Методика определения актуальных угроз безопасности персональных данных при их обработке в информационных системах персональных данных, 2008 год. / Федеральная служба по техническому и экспортному контролю. – URL: https://fstec.ru/tekhnicheskaya-zashchita-informatsii/dokumenty/114-spetsialnye-normativnye-dokumenty/380-metodika-opredeleniya-aktualnykh-ugroz-bezopasnosti-personalnykh-dannykh-pri-ikh-obrabotke-v-informatsionnykh-sistemakh-personalnykh-dannykh-fstek-rossii-2008-god
14. Ворожцова Т.Н. Онтология как основа для разработки интеллектуальной системы обеспечения кибербезопасности // Онтология проектирования. – 2014. – №4. – С. 60 – 76.
15. Ворожцова Т.Н. Разработка онтологии кибербезопасности в энергетике // Information technology and security. – 2013. – №1(3). – С. 19 – 25. – Киев: Институт спец. связи и защиты информации НТУ Украины «КПИ».
16. Montesino R., Fenz S. Automation Possibilities in Information Security Management // Intelligence and Security Informatics Conference (EISIC), 2011 European. – URL: https://www.sba-research.org/wp-content/uploads/publications/PID1947709.pdf (дата запроса: 01.02.17).
17. Birkholz H., Sieverdingbeck I., Sohr K., Bormann C. An Interconnected Asset Ontology in Support of Risk Management Processes, 2012 Seventh International Conference on Availability, Reliability and Security. – URL: http://ieeexplore.ieee.org/document/6329228/ (дата запроса: 01.02.17).
18. Information Security Automation Program/ IBM, 2017. – URL: https://www.ibm.com/support/knowledgecenter/en/SS2TKN_9.1.0/com.ibm.tivoli.tem.doc_9.1/Security_and_Compliance/SCAP_Users_Guide/c_information_security_automatio.html (дата запроса: 04.10.17).
19. Гаськова Д.А., Массель А.Г. Разработка экспертной системы для анализа угроз кибербезопасности в энергетических системах // Информационные и математические технологии в науке и управлении. – 2016. – № 1(27). – С. 116 – 126.
20. Atymtayeva L., Kozhakhmet K., Bortsova G. Building a Knowledge Base for Expert System in Information Security // Soft Computing in Artificial Intelligence. Advances in Intelligent Systems and Computing (№ 270) / eds. Y. Cho, E. Matson– Cham.: Springer, 2014. – URL: http://portal.kazntu.kz/files/publicate/2015-04-20-11866_0.pdf (дата запроса: 12.12.16).
21. Kaur G., Rani P., Garg S. Various Issues in Expert System for Information Management and Audit // International Journal of Advanced Research in Computer Science. – URL: http://www.ijarcs.info/index.php/Ijarcs/article/view/2808 (дата запроса: 01.02.17).
22. Devale A.B., Dr. Kulkarni R.V. A Review of Expert System in Information System audit // International Journal of Computer Science and Information Technologies (IJCSIT). – 2012. – Vol. 3 (5). – P. 5172 – 5175.
23. Adesemowo A.K., Von Solms R., Botha R. A.. IT Asset Ontology for Information Risk in Knowledge Economy and Beyond. – 2016. – Vol. 630. – P. 173.
24. Машкина И.В. Управление защитой информации в сегменте корпоративной информационной системы на основе интеллектуальных технологий: автореф. дис. … д-ра техн. наук. – Уфа, 2009. – 34 с.
25. Рахимов Е.А. Модели и методы поддержки принятия решений в интеллектуальной системе защиты информации: дис. ... канд. техн. наук. – Уфа, 2006. – 237 с. – URL: http://search.rsl.ru/ru/record/01003302763
26. Организация безопасности данных и информационной защиты / НОУ «Интуит». – URL: http://www.intuit.ru/studies/courses/1055/271/lecture/6890?page=5
27. Сухарева С.В., Лыдин С.С., Макейчик Ю.С. Особенности системы централизованного управления и мониторинга средствами защиты информации от несанкционированного доступа // Материалы XVII Международной конференции «Комплексная защита информации. Безопасность информационных технологий». 15-18 мая 2012 г. – Суздаль, 2012. – С. 226 – 228
28. Возможности центра управления сетью (ЦУС) АПКШ «Континент» 3.7 / Код безопасности, 2016. – URL: http://www.securitycode.ru/products/tsentr-upravleniya-setyu-tsus-kontinent-3-7/abilities/
29. Centralized Security Management and Why You Need It / Bitdefender, 2016. – URL: http://businessinsights.bitdefender.com/centralized-security-management
30. СЗИ от НСД «DallasLock» / Центр безопасности информации, 2016. – URL: http://www.cbi-info.ru/groups/page-408.htm
31. Near real-time event management to improve availability and resiliency. TivoliNetcool/OMNIbus / IBM, 2016. – URL: http://www-03.ibm.com/software/ products/ru/ibmtivolinetcoolomnibus
32. Sridhar S., Govindarasu M. Model-Based Attack Detection and Mitigation for Automatic Generation Control // IEEE Transactions on Smart Grid. – URL: http://ieeexplore.ieee.org/document/6740883/ (дата запроса: 02.02.17).
33. Yubin Wang, Wei Wu, Gaofeng Zhan. An Optimized Algorithm in Risk Calculating // 12th International Conference on Computational Intelligence and Security (CIS), (16-19 Dec. 2016), 2017. – URL: http://ieeexplore.ieee.org/document/7820477/ (дата запроса: 02.02.17).
34. Сенцова А.Ю., Машкина И.В. Разработка частной политики информационной безопасности системы облачных вычислений // Вестник Уфимского Государственного авиационного технического университета. – 2016. – №2. – С. 134 – 142.

35. Тулиганова Л.Р., Машкина И.В. Численная оценка риска нарушения информационной безопасности в сегменте виртуализации информационной системы предприятия // Безопасность информационных технологий. – 2015. – №1. – С. 113 – 114.
36. Gribova V.V., Kleshchev A.S., Shalfeeva E.A. Control of Intelligent Systems // J. of Computer and Systems Sciences International. – 2010. – Vol. 49, № 6. – P. 952 – 966.
37. Gribova V., Kleschev A., Krylov D., Moskalenko P., Timchenko V., Shalfeyeva E. A Cloud Computing Platform for Lifecycle Support of Intelligent Multi-agent Internet-services // Intern. Conf. on Power Electronics and Energy Engineering (PEEE). – 2015. – С. 1-7.
38. Для чего нужны обновления антивирусных программ? / Лаборатория Касперского, 2016. – URL: http://www.kaspersky.ru/internet-security-center/internet-safety/antivirus-updates
39. National Vulnerability Database / National Institute of Standarts and Technology. – URL: https://nvd.nist.gov/
40. Банк данных угроз безопасности информации / ФСТЭК России, 2016. – URL: http://www.bdu.fstec.ru/vul
41. Облачная система для обеспечения безопасности веб-трафика Cisco CWS / Cisco, 2016. – URL: http://www.cisco.com/c/ru_ru/products/security/cloud-web-security/index.html
42. Стрельцова С.А. Использование централизованных криминалистических, розыскных, оперативно-справочных и дактилоскопических учетов при расследовании убийств // Сб. трудов конф. «Уголовно-процессуальные и криминалистические средства обеспечения эффективности уголовного судопроизводства». – Иркутск: Байкальский государственный университет, 2014. – С. 343 – 348.
43. Грибова В.В., Краснов Д.А., Островский Г.Е. Медицинские обучающие тренажеры на основе знаний // Материалы 8-й Всероссийской мультиконференции по проблемам управления (МКПУ-2015). 28 сентября – 3 октября 2015 г., с. Дивноморское, Геленджик, Россия. – Ростов-на-Дону: Изд-во Южного Федерального университета, 2015. – С. 48 – 50.
44. Ефимов Б.И., Файзулин Р.Т. Устойчивость объективного решения экспертов при воздействии угроз по блокированию информации в системах принятия решений с привлечение экспертов // Доклады ТУСУР. Томск:ТУСУР. – 2014. – №2(32). – С. 66 – 70.

Материал поступил в редакцию 09.02.18.

Сведения об авторах

ГРИБОВА Валерия Викторовна – доктор технических наук, заместитель директора по научной работе Института автоматики и процессов управления Дальневосточного отделения Российской академии наук (ИАПУ ДВО РАН), г. Владивосток
e-mail:gribova@iacp.dvo.ru

ИВАНОВА Анна Владимировна – аспирант Института автоматики и процессов управления Дальневосточного отделения Российской академии наук (ИАПУ ДВО РАН)
e-mail: 2395146@gmail.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322.2’37 : (048.2)

М.Р. Когаловский

Семантическое аннотирование текстовых документов: основные понятия и таксономический подход*

Одним из инструментов семантического обогащения контента информационных ресурсов является семантическое аннотирование, позволяющее комментировать и оценивать аннотируемые ресурсы и их фрагменты, осуществлять на их основе семантический поиск. Использование таксономического подхода позволяет вместе с тем классифицировать субъекты аннотирования, генерировать новые наукометрические показатели. В статье рассматривается существо семантического аннотирования, определяются основные понятия, обсуждаются общая модель семантической аннотации и таксономический подход к представлению семантики аннотаций, приводятся примеры таксономий, основанных на различных свойствах аннотаций. В качестве примера рассматривается реализация семантического аннотирования в научной информационной системе Соционет.

Ключевые слова: *информационный ресурс, аннотация, семантическая аннотация, общая модель аннотации, таксономия, цитирование, электронная библиотека, система Соционет*

ВВЕДЕНИЕ

В настоящее время активно развивается тенденция повышения семантического уровня представления информационных ресурсов и оперирования ими в информационных системах, прежде всего, применяемых в научных исследованиях. Благодаря этому существенным образом повышается эффективность использования информационных ресурсов, что приобретает особую актуальность в связи с доминантой четвертой парадигмы в научных исследованиях – явления, при котором интенсивное использование данных становится основой и появляется новое направление в информатике – *Data Science*.

Повышение семантического уровня представления информационных ресурсов может быть достигнуто, в частности, путем их семантического обогащения дополнением различного рода сопутствующей информацией. Это делается различными способами. Один из них состоит в использовании *семантического аннотирования* информационных ресурсов, т.е. ассоциирования их с информационными объектами или фрагментами явным образом представленных данных, описывающих семантику этих ресурсов.

Семантическое аннотирование существенным образом обогащает контент научных публикаций. Оно может осуществляться читателем, а в ряде случаев – автоматически с помощью специально разработанного программного обеспечения, способного извлекать необходимую информацию из аннотируемого текста.

Семантические аннотации фрагментов публикаций имеют различный характер. Это могут быть комментарии различного характера к аннотируемому фрагменту текста, оценки его содержания, классификационные рубрики, к которым относится данный фрагмент, какие-либо структурированные метаданные иного рода и т.п. Семантические аннотации используются для разнообразных целей: для обеспечения семантического поиска аннотаций и аннотируемых фрагментов определенных категорий в тексте публикаций; для спецификации оценки содержания аннотируемых фрагментов текста, обогащения их контента путем различного рода дополнений к ним, а также их классификации; для информационного поиска. В электронных библиотеках оценочное семан-

* * Работа выполнена в рамках Государственного задания Институту проблем рынка РАН, тема «Исследование, разработка и поддержка веб-инфраструктуры научных информационных ресурсов с открытым доступом».

тическое аннотирование библиографических ссылок в текстах публикаций на используемые источники позволяет генерировать новые наукометрические показатели.

АННОТИРОВАНИЕ, АННОТАЦИИ И СЕМАНТИЧЕСКИЕ АННОТАЦИИ

В процессе работы с печатным научным текстом читатель часто выписывает цитаты или другие важные для него фрагменты публикации, выделяет их в тексте маркерами, делает комментарии на полях. При чтении текста на компьютере с помощью различных текстовых редакторов все эти возможности также доступны. Так, версии широко распространенного текстового редактора *MS Word* позволяют идентифицировать фрагменты текста шрифтовым выделением или цветом, связывать с нужными фрагментами комментарии. Выделять фрагменты текста цветом и/или сопровождать их комментариями позволяют также продукты компании *Adobe*, такие как *Adobe Reader* или *Adobe Acrobat*, и некоторые другие программные средства. К сожалению, в стандартных вебраузерах при просмотре страниц в формате HTML или XML средства для таких целей не предусмотрены. Для этого нужно использовать другие программные инструменты.

Такое ассоциирование с информационным объектом (например, с фрагментом текста, аудио или видео объектом) различного рода данных, дополняющих и обогащающих тем или иным образом контент, называется его *аннотированием*.

Аннотации как результаты такой работы с текстом или информационным ресурсом другого рода читатель (пользователь) может создавать для собственных целей и/или для других ученых, в том числе, в процессе совместной работы по коллективной подготовке текстового документа или при его экспертизе. В общем случае аннотироваться могут не только тексты, но и информационные ресурсы, представленные в иных средах (графика, аудио, видео).

Аннотирование может осуществляться *в двух формах*. Первая заключается в дополнении к свойствам аннотируемого объекта новых атрибутов, характеризующих некоторые его свойства, которые ранее не были определены. Пример – цветовое или шрифтовое выделение фрагментов текста. В этом случае автономно материализованной аннотации не создается. Вторая форма аннотирования состоит в создании нового информационного объекта, ассоциируемого с аннотируемым (субъектом аннотирования) и отражающего дополнительную информацию (например, комментарий, характеризующий эмоции читателя), оценку его содержания и его класс в некоторой принятой системе классификации (например, с помощью тегов музыкальных клипов, фотографий в коллекции или статей в Википедии), а также различного рода дополнения и т.д.

Такие вновь созданные информационные объекты, ассоциируемые с целевыми объектами, многие авторы называют их *аннотациями*. В англоязычной версии Википедии [1] под аннотацией понимаются «метаданные (например, комментарий, пояснение, разметка представления), которые присоединяются к тексту, изображению или другим данным. Часто аннотации ссылаются на некоторую конкретную часть исходных данных».

В настоящее время созданы компьютерные технологии, предназначенные для аннотирования информационных объектов Веба, которые представлены в различных видах – тексты, аудио, видео и др. Однако для пользователей электронных библиотек и других научных информационных систем особый интерес представляет аннотирование *цифровых текстовых документов*. При этом в качестве целевых информационных объектов могут выступать не только такие документы в целом, но и их отдельные фрагменты.

В ряде развитых электронных библиотек, например, в системе Соционет [2], их информационные ресурсы включают как текстовые документы, так и связи, отражающие различные отношения между ними. В таких случаях субъектами аннотирования могут быть не только фрагменты текстовых документов или документы в целом, но и связи между ними.

Сами аннотации могут быть представлены в виде связей между автором аннотации, информация о котором в электронной библиотеке есть в его персональном профиле, и аннотируемым целевым объектом. В таком случае семантика аннотации может быть представлена семантикой этой связи.

Аннотация целевого объекта может иметь различную семантику, которая представляется явным или неявным образом. Аннотация, представляемая явным образом в терминах какой-либо системы знаний, например, микроформатов, таксономий или онтологий, называется *семантической*. Соответственно, деятельность, продуктом которой являются такие аннотации, естественно называть *семантическим аннотированием.* Назначение семантической аннотации – специфицировать смысл и некоторые свойства аннотируемого информационного ресурса общепонятным образом.

В контексте настоящей работы интерес представляет статья "*What are Semantic Annotations?*" [3], в которой предлагается общий взгляд на аннотирование и некоторая полезная систематизация этой сферы. Предложения авторов статьи базируются на анализе различных подходов к аннотированию ресурсов на примере таких информационных сред, как сематические *Wiki* и семантические блоги, а также систем, использующих *Tagging*. При этом аннотирование рассматривается как присоединение определенных данных к некоторой другой порции данных с установлением того или иного отношения между аннотированными и аннотирующими данными, и предлагается различать три типа аннотаций – неформальные, формальные и онтологические. *Неформальная аннотация* составлена на неформальном языке и поэтому не является *машинно-интерпретируемой* (у авторов машиночитаемой). Напротив, *формальная аннотация* составлена на формальном языке и благодаря этому *машинно-понимаема*. Однако в ней не используются термины онтологии. Наконец, *онтологическая аннотация* (которая, на наш взгляд, является частным случаем формальной аннотации; авторы, вероятно,

понимают ее как семантическую) основана на использовании только терминов онтологии, и поэтому она имеет общепонятный смысл в сообществе, разделяющем эту онтологию.

Неформальная аннотация может быть представлена, например, неструктурированными метаданными, в частности, в виде комментария-пояснения на естественном языке, а формальная – с помощью структурированных метаданных, связывая аннотированный ресурс с некоторой семантической структурой (системой знаний) конкретной предметной области, например, с микроформатами или с онтологией предметной области коллекции текстовых документов.

При использовании онтологии для аннотирования используются ее классы и отношения. В случае использования онтологии для формального описания семантики аннотации (онтологических аннотаций) аннотирование называют *онтологическим*. В более простом случае в качестве аннотирующих данных используются классы таксономии. Такой подход будем называть *таксономическим.*

Наряду с указанными видами аннотаций могут использоваться и *комбинированные аннотации*, состоящие из формального и неформального компонентов. Например, аннотация может указывать класс таксономии, характеризующий свойство аннотируемого объекта, а также содержать текстовый комментарий на естественном языке, выполняющий аналогичную функцию или характеризующий отношение автора аннотации к целевому объекту.

Использование семантического аннотирования существенно обогащает восприятие информационных ресурсов, помогает интерпретировать контент пользователям и механизмам систем, оперирующих с ними, а также обеспечивает дополнительные возможности для большей полноты и точности поиска информационных ресурсов, для их анализа и обработки. На основе коллекций аннотированных научных публикаций семантические аннотации могут также использоваться для генерации новых наукометрических показателей.

Семантическое аннотирование может выполняться *вручную* экспертами, но бывает и *полуавтоматическим* или полностью *автоматическим,* выполняемым с помощью программных систем, которые основаны на извлечении необходимой для этого информации из аннотируемого ресурса. Среди таких систем известны разработки, базирующиеся на наборах данных *Open Linked Data (LOD)* (например, [4] или [5]. Можно упомянуть также завершившийся в 2015 г. проект *Freebase* [6].

Агента (эксперта или программную систему), осуществляющего аннотирование и, следовательно, являющегося автором аннотации далее будем называть *аннотатором*.

В завершение этого раздела необходимо сделать терминологическое замечание. В русскоязычной терминологии в рассматриваемой области существуют два разных термина, позволяющие различать процесс (*аннотирование*) и результат этого процесса (*аннотация*). В то же время в английском языке в обоих этих случаях используется один и тот же термин – *annotation*. Различать его значение в конкретном случае следует в соответствии с контекстом.

ОБЩАЯ МОДЕЛЬ АННОТАЦИИ

Для обсуждения различных подходов к представлению аннотаций, оценки и сравнения функциональных возможностей выразительных средств, используемых для их представления, весьма полезен общий взгляд на феномен аннотации. В уже упоминавшейся выше работе [3] авторы предлагают для этой цели *общую модель аннотации*.

В соответствии с предложенной моделью аннотация рассматривается (в отличие от неформального широко используемого традиционного определения, которое приводят в начале этой работы ее авторы), не просто как некоторые данные, ассоциируемые с аннотируемым информационным ресурсом, а как более сложный объект, состоящий из четырех компонентов:

- *субъект аннотации,*
- ее *объект,*
- *предикат,*
- *контекст аннотации,*

Смысл этих компонентов поясняется в табл. 1.

Каждый из компонентов аннотации может быть формальным или неформальным. Для аннотирования ресурсов Веба понятия формальной и онтологической аннотации определяются в модели более конкретно, причем в качестве критерия формальности компонентов авторы [3] рассматривают их идентифицируемость с использованием стандартных идентификаторов URI.

Если в терминах этой общей модели обратиться к задаче семантического аннотирования, то оно будет заключаться в том, чтобы описать представляющие в данном случае свойства или прокомментировать *объекты и/или предикаты аннотации* средствами какой-либо подходящей системы знаний. Таким образом, общая модель аннотации описывает семантические аннотации как частный случай.

Таблица 1

Компонент аннотации	Пояснение
Субъект аннотации	Аннотируемые данные
Объект аннотации	Аннотирующие данные
Предикат	Отношение между объектом и субъектом аннотации
Контекст	Когда и кем аннотация создана, период времени или область пространства, где она имеет силу и т.п.

ВЫРАЗИТЕЛЬНЫЕ СРЕДСТВА СЕМАНТИЧЕСКОГО АННОТИРОВАНИЯ

По проблематике аннотирования вообще и семантического, в частности, существует обширная литература, посвященная обсуждению различных подходов к аннотированию ресурсов, которые представлены в различных средах и относятся к различным областям приложений. В ней обсуждаются: создание стандартов в этой области, разработки инструментария для автоматизации процесса аннотирования, подходы к семантическому аннотированию на основе различных семантических структур (систем знаний), использование семантического аннотирования в области информационного поиска и извлечения информации из текстов для анализа и обработки аннотированных информационных ресурсов.

Таксономии аннотаций являются одним из продуктивных выразительных средств для описания семантики аннотаций. Рассмотрим несколько представленных в литературе, в том числе, разработанных с участием автора настоящей работы, подходов к описанию семантики аннотаций на основе их классификации с помощью подходящих таксономий, базирующихся на различных свойствах аннотаций. Будем далее называть такое аннотирование *таксономическим аннотированием*. Помимо описания семантики аннотаций, обогащающего контент аннотируемого текста, использование такого подхода позволяет создавать механизмы поиска аннотированных публикаций и фрагментов публикаций, адекватных потребностям пользователей, в частности, ссылок на используемые источники. Кроме того, на этой основе возможна генерация новых нетрадиционных наукометрических показателей. Например, при использовании таксономии, классы которой оценивают аннотируемые фрагменты текста, можно генерировать показатели, которые определяют количество фрагментов данной статьи, оцениваемых позитивно или негативно.

Используемая таксономия обычно зависит от предметной области аннотируемых информационных ресурсов, целей или задач, стоящих перед аннотатором, характера аннотируемых ресурсов (например, фрагменты текста или ссылки на используемые в нем источники), а также от их свойств, существенных в данном конкретном применении.

Рассмотрим ряд известных таксономий аннотаций. Критерии классификации в них – различные свойства объектов и предикатов аннотаций. Прежде всего, обратимся к той же статье [3], в которой предложены заимствованные из ряда других публикаций, называемые авторами этой работы *измерениями,* следующие *критерии* для классификации аннотаций:

- *ассоциация* – способ, при помощи которого аннотация связана с аннотируемым ресурсом. Она может быть как встроенной, так и внешней по отношению к нему и ассоциироваться с ним ссылкой из ресурса;
- *гранулярность* субъекта аннотации – аннотация может относиться к субъекту аннотирования в целом, к какому-либо его разделу или к его другой составной части;
- *особенность представления* – аннотация может относиться к самому документу или к понятиям, описанным в нем либо относящимся к нему;
- *повторное использование терминологии* – аннотация может использовать собственную терминологию или термины из существующих онтологий. Тем самым она является интероперабельной и понятной для других;
- *тип объекта* – объект аннотации может быть литеральным или текстовым, структурным или онтологическим;
- *контекст* – контекст аннотации - когда, кем она создана, в какой сфере, какой срок ее действия и т.п.

Нетрудно видеть, что перечисленные критерии этой классификации представляют собой различные классы «технических» свойств компонентов общей модели аннотаций. Предложенный набор критериев не полон даже относительно этих «технических свойств». Сделать его полным, универсальным на все случаи жизни, конечно, практически невозможно в силу разнообразия потребностей пользователей. Тем не менее, по нашему мнению он полезен для описания не семантики аннотаций, создаваемых в той или иной электронной библиотеке, а, скорее, функциональных возможностей используемого в конкретной системе подхода к аннотированию и/или конкретных инструментов семантического аннотирования, а также для сопоставления функциональности различных подходов/инструментов.

Значимый вклад в создание технологий и инструментария интероперабельного аннотирования, основанного на формальном языке представления аннотаций, внесла деятельность Группы по открытому аннотированию (*Open Annotation Group* или кратко OAG), которая функционирует в последние годы под эгидой консорциума *World Wide Web Center* (W3C). Целью этой группы является разработка спецификаций стандарта онтологии (в терминологии группы используется термин *модель данных*), описываемой на языке RDF, и протокола для открытого интероперабельного аннотирования цифровых документов – текстов, графических изображений, аудио, таблиц и других ресурсов, а также их фрагментов.

В настоящее время предложенные группой спецификации [7-9] приобрели статус рекомендации консорциума и рассматриваются как средство для Семантического Веба, хотя некоторые их элементы могут иметь и более широкое применение. В спецификациях OAG предложена онтология аннотирования, формально определяющая различные виды аннотаций: комментарии, аннотации сущностей (или как теперь принято говорить, вещей), заметок, примеров, опечаток и т.п.

Представляет интерес используемый в онтологии OAG контролируемый словарь мотивов, которыми руководствуется создатель аннотаций (аннотатор). Этот словарь, по существу, может рассматриваться как таксономия мотивов аннотирования, позволяющая явным образом описывать их семантику. Классы словаря мотивов приведены в табл. 2.

В ряде научных электронных библиотек поддерживаются связи между информационными объектами их контента. Частным случаем связей между тек-

стовыми документами являются связи цитирования, представляемые в виде ссылок на используемые или упоминаемые в данной публикации источники, иногда вместе с контекстами этих ссылок. Такие ссылки, как и другие связи, могут рассматриваться как аннотации, а в некоторых случаях – как семантические аннотации. Их компоненты (в терминах общей модели аннотаций) показаны в табл. 3.

Для семантического аннотирования ссылок цитирования также могут использоваться таксономии ссылок. Например, в некоторых случаях целесообразно различать следующие виды ссылок на использованные источники:

- ссылки в тексте цитирующей работы с контекстом – цитатой из цитируемого источника,
- ссылки с иным контекстом,
- ссылки без контекста в цитирующей работе и, наконец,
- ссылки на источники, указанные в ее списке литературы, но с отсутствующими на них ссылками в тексте.

Такая классификация в некоторых случаях может использоваться в качестве таксономии ссылок.

Другой подход к таксономизации ссылок предложен в работе [10], посвященной анализу категоризации влияния цитируемых источников на цитирующие публикации. Предлагается классификация ссылок цитирования в трех измерениях: *функция* (*Function*), *полярность* (*Polarity*) и *влияние* (*Impact*). Для каждого из этих измерений предложен соответствующий набор классов, показанный в табл. 4.

В работе [11] также предлагается классификация ссылок цитирования. Используются иные критерии (табл. 5) по сравнению с рассмотренными выше. Для классификации ссылок по месту в тексте они ранжируются таким образом, что в разделе с результатами их ранг выше, а в обзоре литературы он ниже.

Таблица 2

№ п/п	Мотивация (класс)	Пояснение
1	Оценивание	Аннотация служит для оценки целевого ресурса.
2	Установка закладки	Аннотация отмечает некоторое указанное ее автором место в тексте целевого ресурса.
3	Классифицирование	Аннотация используется для классификации целевого ресурса.
4	Комментирование	Аннотация представляет собой комментарий, относящийся к целевому ресурсу.
5	Описание	Аннотация служит для описания свойств целевого ресурса.
6	Редактирование	Аннотация указывает необходимость редактирования целевого ресурса, например, с тем чтобы устранить опечатку.
7	Выделение маркером	Аннотация указывает намерение ее автора выделить цветом целевой ресурс или его фрагмент для того, чтобы по какой-то причине обратить на него внимание.
8	Идентификация	Аннотация служит для придания индивидуальности целевому ресурсу путем ассоциирования с ним какого-либо уникального идентификатора, например, URI.
9	Связывание	Аннотация определяет связь с некоторым ресурсом, имеющим отношение к целевому.
10	Модерирование	Аннотация служит для указания ценности или качества целевого ресурса, например, для модерирования дискуссий и обсуждений.
11	Запрашивание	Аннотация содержит вопрос о целевом ресурсе.
12	Ответ	В аннотации приводится отклик на целевой ресурс.
13	Создание пометы	Аннотация содержит помету для целевого ресурса.

Таблица 3

Компонент общей модели аннотации	Компонент связи цитирования
Субъект	Цитируемый источник
Объект	Ссылка в цитирующей публикации /Контекст ссылки
Предикат	Класс таксономии
Контекст	Набор атрибутов описателя объекта

Таблица 4

Измерение	Классы
Функция	Цитируемый источник полезен (*Useful*), отражает противоположную точку зрения (*Contrast*), обладает недостатками (*Weakness)*, вносит поправки *(Correct*), уклоняется (*Hedges*), выражает благодарность (*Acknowledge)*, является подтверждением (*Corroboration)*, полемизирует *(Debate*)
Полярность	Позитивная (*Positive*), негативная (*Negative*) и нейтральная (*Neutral*)
Влияние	Негативное (*Negative)*, незначительное (*Perfunctory*) и существенное (*Significant)*

Таблица 5

Критерий	Классы таксономии
Раздел	Абстракт, введение, обзор литературы, методология, результаты/обсуждение, заключение
Интенсивность ссылки	Количество вхождений в тексте ссылки на данный источник
Стиль ссылки	Неконкретное упоминание, конкретное и интерпретирующее упоминание, прямая цитата

В используемых в настоящее время описаниях ссылок цитирования отсутствуют атрибуты, которые бы позволили отобразить их классификацию по приведенным в данной таблице критериям значимости (место в тексте), интенсивности (частотности) и по стилю.

Для того, чтобы их специфицировать, достаточно ввести в таксономию ссылок цитирования три контролируемых словаря:

- *словарь рангов:* высокий, средний, низкий. Этот словарь следует использовать для характеристики значимости ссылки в зависимости от раздела текста, в котором она встречается;
- *словарь интенсивностей*: высокая, средняя, низкая. Его следует использовать для характеристики значимости ссылки в зависимости от частоты ее вхождения в текст;
- *словарь стилей* (характер контекста): прямая цитата, неконкретное упоминание источника, упоминание с пояснением, ссылка без контекста (для случая ссылки в списке литературы, не упоминаемой в тексте).

На основе приведенной классификации ссылок цитирования с помощью указанных контролируемых словарей могут генерироваться новые наукометрические показатели, например: *количество ссылок высокой* (а также средней/низкой) значимости на данную работу, *количество ссылок с высокой* (а также со средней/низкой) интенсивностью, *количество ссылок с прямым цитированием* (а также с интерпретацией в контексте/с неконкретным контекстом/без контекста).

Необходимо упомянуть также онтологию ссылок цитирования C4O (*the Citation Counting and Context Characterization Ontology*) [12], представляющую собой составную часть модульного комплекса онтологий *SPAR* [13], некоторые элементы которых ранее были использованы в таксономии системы Соционет. Онтология C4O включает важные классы отношений между источниками из списков литературы и ссылками на них в текстах публикаций.

Таксономический подход для описания семантики аннотаций используется и в системе Соционет. В ней поддерживается встроенная таксономия, которая подробно описана в работе [14]. Эта таксономия позволяет классифицировать и, тем самым, описывать семантику связей между информационными объектами контента системы.

Некоторые контролируемые словари, составляющие эту таксономию, используются в системе и для семантического аннотирования. В частности, для этой цели предлагается применять *оценочный* контролируемый словарь. Он поможет не только при аннотировании полного текста публикации и ее фрагментов, но и при ссылках на использованные источники в тексте публикации, а также в послестатейном списке литературы. Во всех указанных случаях, кроме последнего, аннотирование может осуществлять любой авторизованный пользователь системы, в последнем случае – только автор данной публикации.

Оценочный контролируемый словарь включает, в частности, следующие классы: *наилучшая, наиболее релевантная работа по обсуждаемой в ней теме; новаторская работа* (результат); *интересная работа* (результат); *оценивается позитивно; оценивается негативно; основывается на заблуждении; возможно, является плагиатом.*

Встроенная в систему Соционет таксономия может легко расширяться путем включения в нее различных контролируемых словарей, позволяющих описывать дополнительные аспекты семантики аннотаций. В настоящее время на стадии обсуждения находится вопрос о дополнении таксономии рядом новых словарей.

Например, для аннотирования фрагментов авторефератов диссертаций и полных текстов диссертаций может потребоваться словарь *научных характеристик*, позволяющий идентифицировать в текстах этих документов важные для их оценки оппонентами фрагменты, содержащие аргументацию соответствия диссертации требованиям ВАК. Словарь включает классы: *актуальность, новизна, достоверность, практическая ценность, теоретическая ценность.*

Полезен также контролируемый словарь, позволяющий специфицировать *статус* аннотируемых фрагментов полного текста публикации: *аксиома*, *доказанное утверждение* (теорема), *цитата из ис-*

пользуемого источника, *фактография*, *результат исследования, постановка задачи.*

Наконец, может быть расширен оценочный словарь путем включения в него следующих дополнительных классов: *актуальная тема исследования, актуальный результат, оригинальный результат, уже известный в науке результат, новый научный результат, фундаментальный результат, обоснованное утверждение, необоснованное утверждение, «вода», раскавыченная цитата.*

Рассмотренные таксономии следует использовать в соответствии с характером аннотируемых ресурсов и с целями (задачами) аннотатора.

СЕМАНТИЧЕСКОЕ АННОТИРОВАНИЕ В СОЦИОНЕТ

В качестве примера применения таксономического подхода к семантическому аннотированию рассмотрим его реализацию в научной информационной системе Соционет. Использованный в системе подход реализован и продолжает развиваться при участии автора данной статьи. Нужно отметить, что таксономия, используемая в системе, ориентирована на описание семантики связей между информационными объектами научной электронной библиотеки и, в частности, семантики аннотаций для научных публикаций.

Система Соционет предусматривает открытое семантическое аннотирование. Это означает, что его возможности доступны любому пользователю системы. Однако необходимо, чтобы он был зарегистрированным и авторизовался при входе в систему, поскольку предусматривается фиксация авторства созданных аннотаций.

Важно отметить, что семантические аннотации в Соционет представляются в форме семантических связей между информационными объектами ее контента. Поэтому в качестве выразительных средств семантического аннотирования в системе используются описания семантических связей. Для их спецификации применяется таксономия семантических связей.

В Соционет поддерживаются информационные объекты – научные публикации, отчеты и произведения других видов, классификаторы и рубрикаторы, в том числе и общепринятые, и семантические связи между ними [15]. Семантика связей определяется с помощью встроенной в систему таксономии, состоящей из нескольких контролируемых словарей. Эта таксономия подробно рассмотрена в работе [14] и кратко обсуждалась в предыдущем разделе вместе с ее возможными расширениями. Классы некоторых контролируемых словарей таксономии используются для описания семантики аннотаций. Это – естественный подход, поскольку аннотации представляются в системе в виде семантических связей.

С точки зрения рассмотренной выше общей модели аннотаций, предложенной в [3], модель аннотаций, используемую в Соционет, можно назвать *комбинированной* – объект аннотации включает формальный и неформальный компоненты. Формальный компонент – это структурированные метаданные, указывающие один из классов подходящего контролируемого словаря встроенной в систему таксономии, определяющий семантику аннотации. Неформальный компонент, называемый в описании аннотации комментарием, - это неструктурированные метаданные, представленные в виде текста на естественном языке.

Субъектами аннотирования в Соционет могут быть полные тексты представленных в системе публикаций, фрагменты их абстрактов, а также фрагменты полных текстов. Кроме того, аннотироваться могут связи цитирования одних публикаций в других. Связи этого вида – это ссылки на источники из послестатейных списков литературы в текстах, а также сами библиографические описания использованных источников в этих списках.

Наряду со связями цитирования, выделяемыми пользователем-аннотатором в «ручном режиме», субъекты аннотирования такого рода могут порождаться в автоматическом режиме средствами анализа полных текстов публикаций, представленных в контенте системы в PDF-формате.

Соционет является мультипользовательской системой, и поэтому для одного субъекта аннотирования может быть создано несколько аннотаций одним или разными пользователями. Аннотации представляются в Соционет в виде классифицированных связей «персона – субъект». Их описания включают идентификацию персоны-автора аннотации, идентификацию субъекта аннотации, класс выбранного аннотатором контролируемого словаря таксономии, а также текстовый комментарий.

Функциональные возможности Соционет позволяют использовать ее как платформу для виртуальной коммуникационной среды научного сообщества пользователей [16, 17]. Эти возможности основаны на реакциях авторов публикаций на появление семантических связей этих публикаций с публикациями других авторов либо оценочных связей, касающихся этих публикаций. Такая реакция состоит в создании новой связи профиля ее автора со связью, на появление которой он реагирует.

Поскольку аннотации представляются в виде семантических связей, указанные возможности могут быть применены и к ним. Поэтому создание аннотаций потенциально может быть вовлечено в процессы коммуникаций.

Формальные компоненты объектов аннотаций в Соционет – структурированные метаданные – могут использоваться в критериях поиска аннотаций, интересующих пользователя классов, а также для генерации ряда новых наукометрических показателей наряду с другими, автоматически формируемыми сервисами системы.

Для возможности создания новых наукометрических показателей в описания создаваемых связей должны быть перенесены классификационные атрибуты цитирования. В Соционет также необходимо создать соответствующие сервисы, которые будут генерировать и показывать полученные показатели на странице метаданных (описателя) публикации, как это реализовано сегодня для других показателей в системе.

Этими новыми показателями могут быть следующие: *количество ссылок высокой* (а также средней/низкой) значимости на данную работу, *количество ссылок с высокой* (а также со средней/низкой интенсивностью), *количество ссылок с прямым ци-*

тированием (а также с интерпретацией в контексте/с неконкретным контекстом/без контекста).

Следует отметить, что механизмы поддержки таксономии семантических связей обеспечивают ее расширение путем задания новых или дополнения составляющих ее контролируемых словарей и настройки соответствующих компонентов программного обеспечения системы. Тем самым система обладает расширяемостью выразительных средств семантического аннотирования.

ЗАКЛЮЧЕНИЕ

Семантическое аннотирование научных публикаций позволяет существенно обогащать их контент. Таксономии аннотаций являются достаточно содержательным, простым для использования и программной реализации выразительным средством представления семантических аннотаций. Структурированный характер последних в таком представлении обеспечивает легкое восприятие пользователями и позволяет осуществлять на их основе информационный поиск аннотированных текстовых документов в коллекциях электронных библиотек и отдельных фрагментов документов, генерировать новые нетрадиционные наукомерические показатели.

СПИСОК ЛИТЕРАТУРЫ

1. Annotation. Wikipedia. – URL: https://en.wikipedia.org/wiki/Annotation (дата обращения: 20.12.2017).
2. Parinov S., Lyapunov V., Puzyrev R., Kogalovsky M. Semantically Enrichable Research Information System SocioNet // Knowledge Engineering and Semantic Web. 6th Intern. Conf. KESW 2015 / eds. P. Klinov, D. Mouromtsev. – NY: Springer, 2015. – P. 147-157 (The Communications in Computer and Information Science series Vol. 518). DOI: 10.1007/978-3-319-25543-0_11
3. Oren E., Hinnerk Moller K., Scerri S., Handschuh S., Sintek M. What are Semantic Annotations? – URL: http://www.siegfried-handschuh.net/pub/2006/whatis semannot2006.pdf (дата обращения: 20.12.2017).
4. Gagnon M., Zouaq A., Jean-Louis L. Can we use Linked Data Semantic Annotators for the Extraction of Domain-Relevant Expression. In: WWW 2013 Companion, pp. 1249-1246 (2013).
5. DBpedia. Википедия. – URL: https://ru.wikipedia.org/wiki/DBpedia (дата обращения: 20.12.2017).
6. Freebase. In: Wikipedia. – URL: https://en.wikipedia.org/wiki/Freebase (дата обращения: 20.12.2017).
7. Web Annotation Data Model. W3C Recommendation 23 February 2017. – URL: https://www.w3.org/TR/2017/REC-annotation-model-20170223/ (дата обращения: 20.12.2017).
8. Web Annotation Protocol. W3C Recommendation 23 February 2017. – URL: http://www.w3.org/TR/annotation-protocol/ (дата обращения: 20.12.2017).
9. Web Annotation Vocabulary. W3C Recommendation 23 February 2017. – URL: http://www.w3.org/TR/annotation-vocab/ (дата обращения: 20.12.2017).
10. Hernández-Alvarez M., Gómez Soriano J.M., Martinez-Barco P. Citation function, polarity and influence classification. DOI: 10.1017/S1351324916000346 (2017).
11. Zhang G., Ding Y., Milojević S. Citation Content Analysis (CCA): A Framework for Syntactic and Semantic Analysis of Citation Content. arXiv:1211.6321 (2012).
12. Shotton D. C40, the Citation Counting and Context Characterization Ontology. Version 1.1.1, 11/05/2013. – URL: http://purl.org/spar/c4o (дата обращения: 20.12.2017).
13. SPAR Ontologies. Describing Publishing Domain. – URL: http://purl.org/spar/ (дата обращения: 20.12.2017).
14. Когаловский М.Р., Паринов С.И. Таксономия семантических связей информационных объектов контента научной электронной библиотеки // Научно-техническая информация. Сер. 2. – 2015. – № 9. – С. 15-23; Kogalovskii M.R., Parinov S.I. The Taxonomy of Semantic Linkages of Information Objects in Research Digital Library Content // Automatic Documentation and Mathematical Linguistics. – 2015. – Vol. 49, № 5. – P. 163-171. DOI: 10.3103/S0005105515050027
15. Паринов С.И., Когаловский М.Р. Технология семантического структурирования контента научных электронных библиотек. Труды XIII Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL-2011. Воронеж, 19-22 октября 2011 г.». – Воронеж: Воронежский гос. ун-т, 2011. – С. 197-206. DOI: 10.13140/2.1.2150.7525
16. Kogalovsky M.R., Parinov S.I. Scholarly Communications in a Semantically Enrichable Research Information System with Embedded Taxonomy of Scientific Relationships // Knowledge Engineering and Semantic Web. 6th International Conference KESW 2015 / eds. P. Klinov, D. Mouromtsev. – NY: Springer, 2015. – P. 87-101 (The Communications in Computer and Information Science series Vol. 518). DOI: 10.1007/978-3-319-24543-0_7
17. Когаловский М.Р., Паринов С.И. Виртуальная научная коммуникационная среда на основе семантической научной информационной системы // Научно-техническая информация. – 2016. – № 8. – С. 19-25; Kogalovskii M.R., Parinov S.I. A Virtual Scientific Communication Environment Based on a Semantic Scientific Information System // Automatic Documentation and Mathematical Linguistics. – 2016. – Vol. 50, № 5. – P. 189-194. DOI: 10.3103/S0005105516050046

Материал поступил в редакцию 08.01.18.

Сведения об авторе

КОГАЛОВСКИЙ Михаил Рувимович – кандидат технических наук, доцент, зав. лабораторией Института проблем рынка РАН, Москва
e-mail: kogalov@gmail.com

УДК 025.45

В. Н. Белоозеров

Правила алгоритмического порождения индексов УДК для тематической классификации информационных ресурсов*

Изложены формальные правила перечисления всех возможных индексов Универсальной десятичной классификации (УДК), которые уточняют рекомендации, приводимые в руководствах по использованию УДК для тематического индексирования научно-технических документов. Правила составляют алгоритм развёртывания сложной структуры индексов УДК по непосредственным составляющим с последующим редактированием по ряду трансформационных операций. Алгоритм позволяет установить соответствие структуры любого индекса правилам за конечное число шагов.

Ключевые слова: *Универсальная десятичная классификация, индексы УДК, порождающая грамматика, предметное индексирование документов*

ВВЕДЕНИЕ

Универсальная десятичная классификация (УДК) [1] является стандартным инструментом для описания содержания научных работ в информационно-библиотечной практике и автоматизированных системах научно-технической информации. Рядом национальных стандартов России [2, 3] применение УДК признано как обязательное для всей издательской продукции и электронных информационных ресурсов.

Индексы УДК отражают содержание документа условными кодами. Они образуются согласно довольно сложной системе правил, которые излагаются в описаниях классификационных таблиц. Разъяснению правил посвящена довольно обширная литература (см., напр. [4, 5]). При интуитивной ясности этих правил на интеллектуальном уровне, они недостаточно чётко формализованы, что затрудняет применение к ним автоматических методов анализа, а в ряде случаев ведёт к неоднозначной интерпретации.

В задачах автоматической обработки индексов УДК необходимо иметь способ определения корректности анализируемых выражений. В настоящей статье изложены правила перечисления правильных индексов УДК, которые дают принципиальную возможность проверки корректности каждого индекса. Предварительная версия изложенных здесь правил доложена на конференции [6].

Алгоритм перечисления индексов УДК состоит из двух принципиально разных частей. Сначала мы строим множество $\boldsymbol{U}$ «правильных формул». В него входят цепочки цифр, специальных знаков и букв, которые отличаются от правильных индексов УДК тем, что цифры не разделены «зрительными точками» и не сделаны некоторые возможные сокращения длины кодов. Затем операции «редактирования формул» приводят к множеству правильных индексов $\boldsymbol{W}$. Это множество – бесконечное счётное, перечисление его элементов по этим правилам идёт от индексов простой структуры к сложным и составным индексам.

Алгоритм формирования множества правильных формул $\boldsymbol{U}$ состоит в циклическом выполнении операций по правилам 1–4.7. Для получения множества $\boldsymbol{W}$ правильных индексов УДК на любом шаге формирования множества $\boldsymbol{U}$ к его элементам применяются правила 5.1 – 5.4.

Отметим, что данный алгоритм указывает все формально правильные индексы УДК, не принимая во внимание их реальную употребимость. Так, например, из десяти возможных главных классов УДК, обозначаемых одной цифрой в диапазоне от 0 до 9, реально используют только девять, а цифра 4, формально способная представлять некоторый тематический класс, в настоящее время не используется, а зарезервирована для возможного обновления и развития структуры классификационной системы в будущем.

Далее следуют правила, которые излагаются в разных формулировках (за редким исключением), отличающихся степенью формализованности, с тем чтобы их можно было понять как читателю, не искушённому в математической символике, так и читателю, желающему понять точный математический смысл операций.

* Работа выполнена в рамках проекта РФФИ № 17-07-00153 «исследование системы классификаторов по науке и технике и разработка механизма смысловой навигации и поиска знаний в информационных сетях»

1. АЛФАВИТ ИНДЕКСОВ

Все правильные формулы УДК строются (как правило) из символов, образующих два множества – ***A*** и ***S***.

● Множество цифр $A \equiv \{$ **0, 1, 2, 3, 4, 5, 6, 7, 8, 9** $\}$.

● Множество разделителей $S \equiv \{$ **‘ , - , º , = , + , / , (,) , [,] , « , » , *** $\}$.

Иными словами, в ***S*** входят символы: апостроф, дефис, градус, равенство, плюс, косая дробь, процент, круглые и квадратные скобки, кавычки, астериск. Градус сейчас временно представляет символ определителя «точка-ноль», который состоит из двух типографских знаков, но функционирует как единый элемент.

В некоторых случаях (см. правила 3.3, 3.4, 4.7 и 4.8) к индексам УДК присоединяют цепочки символов, включающие **буквы,** которые составляют множество ***L***. В это множество всегда входят все буквы основного латинского алфавита, а для национального применения включают также буквы национальных алфавитов, в частности – русского.

Множество ***A*** играет особую роль в системе УДК: оно задаёт членение всей области классифицирования на десять тематических разделов. Этот факт подчёркивает первое правило формирования индексов.

Правило 1. Во множество правильных формул включается всё множество цифр ***A***. ***Каждый элемент множества A (цифра) является правильной формулой УДК.*** Если $a \in A$, то $a \in U$, $A \subset U$ и $A \subset W$.

Формулы, состоящие из одной цифры, являются одновременно и правильными индексами УДК, которые соответствуют **главным классам** УДК, а именно:

0 Общий отдел
1 Философия. Психология
2 Религия. Богословие
3 Общественные науки
4 (пустой резервный класс)
5 Математика. Естественные науки
6 Прикладные науки. Медицина. Технология
7 Искусство. Развлечения. Зрелища. Спорт
8 Язык. Языкознание. Лингвистика. Литература
9 География. Биографии. История.

2. ПРОСТЫЕ КЛАССЫ

Наиболее обычны для УДК индексы, состоящие только из цифр, не содержащие букв и элементов множества ***S***. Эти индексы, называемые **простыми**, обозначают **простые классы** УДК, которые образуются последовательным делением каждого главного класса на подклассы, число которых не превосходит десяти. Каждый подкласс обозначается индексом своего надкласса с добавлением справа ещё одной цифры – номера подкласса на данной ступени членения классов. При этом после каждой третьей цифры подряд в индексе УДК принято ставить точку для облегчения зрительного восприятия. Эти «зрительные» точки мы временно игнорируем и рассматриваем не «индексы», а «формулы» УДК – цепочки символов, не содержащие точек. Таким образом, любая конечная цепочка цифр может представлять обозначение результата последовательного деления одного из главных классов УДК (номер которого – первая цифра цепочки). Следовательно, любая конечная цепочка является **правильной формулой** УДК.

Этот результат можно сформулировать как следующее правило:

Правило 2. Если $a_i \in A$, при $i \in \{1, 2, \ldots, n\}$, где $n \geq 1$, то $a_1a_2\ldots a_i\ldots a_n \in U$.

Множество всех конечных цепочек цифр обозначим как ***B***. Это множество автоматически включает в себя множество одиночных цифр $A \subset B$.

Правило 2 можно было бы сформулировать и по-другому, как циклическое наращивание цифр на индекс главного класса:

Правило 2а. Если $a \in A$ и $b \in U$, то $a \in U$ и $ba \in U$. $B \subset U$.

Это обстоятельство можно выразить процессуальным правилом последовательного построения (перечисления) элементов множества: к каждому элементу множества ***B*** можно справа присоединить цифру; результат будет также правильной простой формулой, входящей в ***B***. Если ***b*** – элемент множества ***B***, то элементом множества ***B*** является также конкатенация (непосредственное присоединение) ***a*** к ***b***.

Формулы, полученные по правилу 2, будем называть **простыми формулами УДК**. Они представляют собой конечные цепочки цифр и образуют бесконечное счётное множество ***B***. Эти формулы соответствуют индексам простых классов УДК, получаемых последовательным разбиениям главных классов на подклассы, номера которых и составляют последовательность цифр в цепочках формул.

Примеры индексов, соответствующих простым формулам:

00 *Общие вопросы науки и культуры*
001 *Наука и знание в целом. Науковедение. Организация умственного труда*
159.96 *Особые психические состояния и явления*
621.313.8 *Электрические машины с постоянными магнитами*
811.511.111 *Финский язык*

3. УСЛОЖНЁННЫЕ ФОРМУЛЫ

Простые формулы УДК соответствуют классам разбиения универсума знаний на чисто тематические области, но документы и исследования, описываемые индексами УДК, могут быть охарактеризованы по ряду аспектов, не связанных с тематикой, например, по форме документа, языку, географическому месту, времени, материалу. Такие признаки включаются в индекс УДК с помощью «общих» и «специальных» **определителей**. Каждый определитель строится аналогично простым индексам, но начинается с одного из символов, входящих во множество ***S*** (и в ряде случаев заключён в скобки). К одному простому индексу может быть присоединено несколько определителей. Кроме того, некоторые общие определители могут выступать как самостоятельные индексы УДК.

Множество формул УДК, включающее простые и усложнённые формулы, обозначим буквой ***C***. Его формирование можно описать следующими правилами.

Правило 3.1. Все элементы ***B*** включаем в ***C***. Если $b \in B$, то $b \in C$, следовательно $b \in U$ и $A \subset B \subset C \subset U$.

Правило 3.2. Если $c_1 \in C$ и $c_2 \in C$, то множество $\{c_1{'}c_2, c_1\text{-}c_2, c_1{}^{o}c_2, c_1{=}c_2, c_1(c_2), c_1(=c_2), c_1«c_2», {=}c_2, (c_2), (=c_2), «c_2»\}$ входит в C, которое входит в U, и $A \subset B \subset C \subset U$.

В этом правиле учтены все определители: $'c_2, \text{-}c_2, {}^{o}c_2$ – три серии специальных определителей, значение которых задается в каждом разделе УДК индивидуально, общие определители: $\text{-}0c_2$ – свойств, процессов и материала; $=c_2$ – языка изложения документа; $(=c_2)$ – описываемого этноса; (c_2) – формы документа или географического места; $«c_2»$ – времени.

Примеры индексов с определителями:

(470) *Европейская часть Российской Федерации* (географический определитель как самостоятельный индекс, обозначающий территорию).

069 *Музеи. Постоянные выставки*

069.01 *Теория музейного дела. Музееведение* (***.01*** – определитель теоретического метода рассмотрения)

069(4) *Музеи Европы* (***(4)*** – определитель европейской локализации объекта)

069(=72) *Музеи, посвящённые австралийским аборигенам* (***(=72)*** – этнический определитель австралийских аборигенов)

81 *Языкознание и языки. Лингвистика*

81'0 *Происхождение и периоды развития языков* (***'0*** – определитель истории языков)

81=161.1 *Документы по языкознанию на русском языке* (***=161.1*** – определитель русского языка)

В необходимых случаях индексы УДК разрешается уточнять добавлением справа «внесистемных элементов» – кодов других классификаций и имён собственных. Посторонние коды присоединяются через знак астериска *, а имена собственные должны начинаться с буквы, желательно заглавной. Однако на структуру этих добавлений следует ввести ограничение: они не должны содержать элементов множества ***S***; иначе будет происходить ложное отождествление с определителями, имеющими своё особое значение. Формулы, усложнённые внесистемными добавлениями, также включаем во множество ***C*** согласно следующим двум правилам.

Правило 3.3. **Астериск.** К каждой формуле, принадлежащей множеству ***C***, может быть присоединена цепочка символов из множеств ***A*** и ***L***, начинающаяся с астериска. Если $d \in C$ и $\{a_1, a_2, \ldots, a_n\} \subset (A \cup L)$, то $d{*}a_1a_2\ldots a_i\ a_n$ принадлежит ***C*** при $1 \leqslant i \leqslant n$.

Пример: ***630*2*** *Лесоводство* (Членение класса ***630*** *Лесное хозяйство* в УДК заимствовано из специальной лесотехнической классификации)

Правило 3.4. **Алфавитное дополнение.** К каждой формуле, принадлежащей множеству ***C***, может быть присоединена цепочка символов из множеств ***A*** и ***L***, начинающаяся с буквы (элемента множества ***L***). Если $d \in C$ и $\{a_1, a_2, \ldots, a_n\} \subset (A \cup S \cup L)$ и $b \in L$, то $dba_1a_2\ldots a_i\ a_n \in C$ при $i \leqslant n \geqslant 1$.

Пример: ***8211611A.С.Пушкин*** – формула, соответствующая индексу ***821.161.1A.С.Пушкин***, который имеет значение «произведения А. С. Пушкина как явление русской литературы».

Формулы, образованные по правилам 3.2–3.4, будем называть **усложнёнными формулами УДК.** Они соответствуют простым индексам УДК с добавлением специальных и общих определителей и «внесистемных» расширителей (алфавитных расширений и заимствованных кодов с астериском).

4. ФОРМИРОВАНИЕ «СОСТАВНОГО» МНОЖЕСТВА ФОРМУЛ D

Во-первых, во множество ***D*** включаем все простые формулы по правилу 4.1, но главное содержание ***D*** принадлежит индексам, отражающим сложную тематическую структуру документа.

Правило 4.1. Все элементы множества ***C*** включаем во множество ***D***. Если $c \in C$, то $c \in D$.

Документы, включающие рассмотрение нескольких тем или обсуждающие тему под разными аспектами, могут относиться к двум и более разным классам УДК. Индексы таких документов формируются путём комбинации индексов, отражающих включённые в документ темы и аспекты. Если документ охватывает две темы в их относительной полноте, то индексы этих тем соединяются знаком «плюс» +, и такому составному индексу соответствует объединение тем. Если в документе рассматриваются только вопросы, связанные со взаимодействием тем, с явлениями, относящимися сразу к двум темам, то индексы соединяют знаком двоеточия, и в данном случае происходит пересечение тематики соединяемых классов. В том случае, когда тематика одного из классов является лишь отличительной чертой в рассматриваемом явлении и не обсуждается в документе по существу, индекс такого класса присоединяют к индексу основной тематики двойным двоеточием. Формулы, отражающие такие индексы, образуются по следующему правилу.

Правило 4.2. Пары элементов множества ***D***, соединённые знаками плюс, двоеточие и двойное двоеточие, включаем во множество ***D***. Если d_1 и d_2 принадлежат множеству ***D***, то множеству ***D*** принадлежат также выражения $d_1{+}d_2, d_1{:}d_2, d_1{::}d_2$. Если d_1 и $d_2 \in C$, то $\{d_1{+}d_2, d_1{:}d_2, d_1{::}d_2.\} \subset D$.

Примеры составных индексов:

69+72 *Строительство и архитектура*

378:005 *Менаджмент в высшем образовании.*

504.61::355 *Ущерб среде от военной деятельности* (Здесь ***504.61*** – ущерб среде от деятельности человека, ***355*** – военное дело).

Если тематика документа охватывает несколько классов УДК, расположенных в классификационной таблице рядом, то индекс этого документа образуется путём соединения индекса начального класса диапазона с индексом конечного класса при помощи знака косой дроби в соответствии со следующим правилом.

Правило 4.3. Во множество ***D*** включаем пары простых формул из ***B***, соединённые знаком косой дроби /, если алфавитный порядок первой формулы

из соединяемой пары предшествует второй формуле. Если b_1 и b_2 принадлежат множеству B, и десятичная дробь $0,b_1$ меньше дроби $0,b_2$, $0,b_1<0,b_2$, то выражение b_1/b_2 принадлежит множеству D. Если $b_1 \in B$, $b_2 \in B$ и $0,b_1<0,b_2$, то $b_1/b_2 \in D$.

Примеры «диапазонных» индексов:

21/29 *Религиозные системы. Религии, верования* (Включает все классы региональных, исторических и мировых религий).

551.1/551.4 *Общая геология. Физическая и динамическая геология*

Если документ описывается тремя или более классами УДК, то в комбинированном индексе путём расстановки скобок можно указать характер и направление взаимодействий соответствующих аспектов содержания. Компоненты составного индекса, описывающие тесно взаимодействующие темы, заключают в квадратные скобки. Это следующее правило построения составных формул.

Правило 4.4. Если d_1, d_2 и d_3 принадлежат D, то включаем в D также формулы $[d_1 \clubsuit d_2] \clubsuit d_3$ и $d_1 \clubsuit [d_2 \clubsuit d_3]$, где $\clubsuit$ – любой из символов {+ , : , :: } (плюс, двоеточие, двойное двоеточие).

Составные индексы могут быть, в свою очередь, усложнены общими определителями (которые также могут описываться составными формулами) и алфавитным расширением. Правило, относящееся к общим определителям.

Правило 4.5. Если d_1, d_2 и d_2 принадлежат D, то множество комбинаций $\{[d_1 \clubsuit d_3]=d_2$, $[d_1 \clubsuit d_3](d_2)$, $[d_1 \clubsuit d_3](=d_2)$, $[d_1 \clubsuit d_3]«d_2»$, $=[d_1 \clubsuit d_3]$, $(d_1 \clubsuit d_3)$, $(=[d_1 \clubsuit d_3])$, $«d_1 \clubsuit d_3»\}$ также входит в D.

Аналогично правилу 3.4 составные индексы могут быть взяты в скобки и дополнены строкой символов, начинающихся с буквы.

Правило 4.6. Алфавитное дополнение. (Аналогично 3.4) К каждой составной формуле, принадлежащей множеству D, может быть присоединена цепочка символов из множеств A и L, начинающаяся с буквы (элемента множества L). Если $d_1 \in D$, $d_2 \in D$, $\{a_1, a_2, \ldots, a_n\} \subset (A \cup L)$ и $b \in L$, то $[d_1 \clubsuit d_2]ba_1a_2 \ldots a_i \ldots a_n \in D$ при $1 \leqslant i \leqslant n$.

Формулы, образованные по правилам 4.4–4.6, будем называть **скобочными формулами** УДК. Скобочные формулы в совокупности с формулами, построенными по правилам 4.2 и 4.3, будем называть **составными формулами УДК.**

Пример скобочной формулы:

[336146::0049]:[5508+6221/6222].

Расстановка точек после каждой третьей цифры переводит эту формулу в правильный индекс ***[336.146::004.9]:[550.8+622.1/622.2],*** имеющий значение «***компьютерная проверка бюджетных расходов по поиску, разведке и разработке месторождений***», которое складывается из значений составляющих индексов: ***336.146*** *– проверка бюджетных расходов,* ***004.9*** *– прикладные компьютерные технологии,* ***550.8*** *– поиск и разведка месторождений,* ***622.1*** *– доразведка месторождений,* ***622.2*** *– разработка месторождений.*

Правило 4.7. Составные формулы включаем во множество U. $D \subset U$.

Составные формулы соответствуют индексам, представляющим объединения и пересечения классов, обозначаемых простыми и усложнёнными формулами. На этом исчерпываются модели построения формул УДК. Остаётся перевести их в правильные индексы путём расстановки точек и сделать ещё некоторые (факультативные) сокращения.

5. РЕДАКТИРОВАНИЕ ФОРМУЛ

Построение множества W правильных индексов УДК осуществляется путём расстановки точек в элементах множества формул U по правилам 5.1 и 5.2.

Правило 5.1. Заменяем символ градуса последовательностью точка-нуль. Если $v \in U$ и $v \equiv e_1e_2 \ldots e_i \ldots e_n$, где e_i – какой-либо символ из $A \cup S$, то $w \in W$, если $w \equiv f_1f_2 \ldots f_i \ldots f_n$, где для всех $i \in \{1,\ldots,n\}$ $f_i \equiv e_i$, если $e_i \neq {}^{o}$, и $f_i \equiv .0$, если $e_i \equiv {}^{o}$.

Правило 5.2. После каждой третьей цифры в последовательности цифр вставляем точку (расстановка зрительных точек).

Если $ab_1b_2b_3 \ldots b_nc \in U$, где a и c – произвольные цепочки символов, а b_i – цифры ($i \in [3, n]$, $n \geq 4$), то в U включаем также $ab_1b_2b_3. \ldots b_nc$. (После b_3 стоит точка!).

Следующие операции редактирования не увеличивают «семантической силы» УДК, т. е. не приводят к построению индекса, имеющего смысловую интерпретацию, отличную от исходного, но позволяют несколько сократить длину записи.

Правило 5.3. Сокращение диапазонов.

Во втором элементе диапазона можно опустить все символы, совпадающие с началом первой части диапазона, вплоть до последней точки в этой цепочке совпадающих символов. При этом сама последняя точка должна остаться в индексе.

Если x_1 и x_2 принадлежат множеству правильных индексов W, $x_1 \equiv a_1 \ldots a_n.b$ и $x_2 \equiv a_1 \ldots a_n.c$, а b и c – какие-либо последовательности элементов множеств A и S, то $x_3 \equiv a_1 \ldots a_n.b/.c$ входит в W и является правильным индексом ***УДК.***

Индексы с сокращёнными диапазонами записываем в W в дополнение к индексам с развёрнутыми диапазонами.

Пример: Индекс класса ***539.3/.6*** *Сопротивление материалов* эквивалентен по значению индексу ***539.3/539.6.***

Правило 5.4. Объединение скобочных определителей. При сочетании двух ***не составных*** скобочных определителей цепочки символов, указанные в первой колонке представленной ниже таблицы, заменяем символами, указанными во второй колонке.

Если цепочки символов ***(a)(b), (a):(b), (a)::(b), (a)+(b), (a)/(b),*** где a и b – произвольные цепочки символов, принадлежат множеству правильных индексов W, то этому множеству принадлежат также цепочки ***(a:b), (a:b), (a::b), (a+b), (a/b)*** – соответственно.

Объединение скобочных определителей

Исходная цепочка символов	Сокращённая цепочка символов	Комментарий
(…)(…)	*(… : …)*	Два определителя со скобками, стоящие рядом, можно объединить в одну скобочную конструкцию, поставив между символами определителей знак отношения (двоеточие).
(…):(…)	*(… : …)*	Два определителя со скобками, соединённые двоеточием, можно объединить в одну скобочную конструкцию, удалив внутренние скобки.
(…)::(…)	*(… :: …)*	Два определителя со скобками, соединённые двойным двоеточием, можно объединить в одну скобочную конструкцию, удалив внутренние скобки.
(…)+(…)	*(… + …)*	Два определителя со скобками, соединённые знаком плюс, можно объединить в одну скобочную конструкцию, удалив внутренние скобки.
(…)/(…)	*(… / …)*	Два определителя со скобками, соединённые знаком дроби, можно объединить в одну скобочную конструкцию, удалив внутренние скобки.

Пример: Индекс класса ***(211/213)*** *Климатические пояса* эквивалентен индексу ***(211)/(213).***

Правило 5.5. Удаление внешних скобок. Квадратные скобки, стоящие в начале и в конце правильного индекса УДК, можно удалить, не нарушая этим правильности и смысла индекса. Если ***[a1*** и ***a2]*** принадлежат ***W***, где ***a1*** и ***a2*** – произвольные цепочки символов из ***A***, ***S*** и ***L***, то ***a1*** и ***a2*** также принадлежат ***W***.

Пример:

Рассмотренный ранее индекс

[336.146::004.9]: [550.8+622.1/622.2]

может быть записан в виде:

336.146::004.9]:[550.8+622.1/622.2.

Формулы, преобразованные по правилам 5.1–5.5, записываем во множество правильных индексов ***W.***

ЗАКЛЮЧЕНИЕ

Последовательное выполнение приведенных здесь правил (в произвольном порядке) приводит к постепенному расширению множества формально правильных индексов УДК, которое включает:

- десять цифр {0, 1, 2, 3, 4, 5, 6, 7, 8, 9} – индексов **главных классов УДК**;
- бесконечное счётное множество всех цепочек цифр, разделённых точками на тройки – **простые индексы УДК**;
- бесконечное счётное множество **усложнённых индексов УДК**, представляющих собой простые индексы с присоединёнными к ним посредством символов определителей дополнительных цепочек цифр, а также цепочек алфавитно-цифровых символов, присоединённых посредством астериска или произвольной буквы;
- бесконечное счётное множество **составных индексов УДК**, представляющих собой простые и усложнённые индексы, объединённые в цепочки символов знаками присоединения, отношения и квадратными скобками.

Любой правильный индекс УДК может быть построен по этим правилам за конечное число шагов[*].

На основе изложенных здесь правил построения индексов УДК может быть разработан алгоритм их анализа и интерпретации путём составления осмысленного текста из формулировок наименований классов УДК, коды которых составляют анализируемый индекс.

СПИСОК ЛИТЕРАТУРЫ

1. УДК. Универсальная десятичная классификация. – 4-е полн. изд. на русск. яз.: в 10 т., 12 кн. – М.: ВИНИТИ. 2001 – 2011.
2. ГОСТ Р 7.0.59–2008 Система стандартов по информации, библиотечному и издательскому делу. Индексирование документов. Общие требования к систематизации и предметизации.
3. ГОСТ 7.90–2007 Система стандартов по информации, библиотечному и издательскому делу. Универсальная десятичная классификация. Структура, правила ведения и индексирования.
4. McIlwaine I.C. The Universal Decimal Classification: A guide to its use. – The Hague: UDC Consortium, 2007. – 278 p.
5. Антошкова О.А., Астахова Т.С., Белоозеров В.Н. и др. Учебное пособие по Универсальной десятичной классификации. – 3-е изд. испр. и доп. – М.: ВИНИТИ, 2014. – 186 с.
6. Белоозеров В.Н. Алгоритм построения индексов УДК // Материалы научно-практической конференции «Перспективные направления научных исследований и критические технологии в классификационных системах» (Москва, 25-27 октября 2017 г. – М.: ВИНИТИ, 2017. – С. 36–39.

Материал поступил в редакцию 20.02.18.

Сведения об авторе

БЕЛООЗЕРОВ Виктор Николаевич – кандидат филологических наук, доцент, ведущий научный сотрудник ВИНИТИ РАН, Москва
e-mail: systemling@narod.ru

[*] Здесь не рассматривается трансформация «интеркаляции», при которой общий определитель вставляется в середину цифровой цепочки. Эта операция иногда применяется для формирования надлежащего полочного индекса в библиотечных хранилищах, но при первичном индексировании документов интеркаляция неуместна.

Герольд Георгиевич Белоногов
(05.06.1925 – 18.04.2018)

Ушел из жизни доктор технических наук, профессор Герольд Георгиевич Белоногов, крупный деятель в области информатики, компьютерной лингвистики, автоматической обработки текстов. С ним ушла целая эпоха становления и развития отечественной науки в этой области.

В 1947 г. Г.Г. Белоногов окончил Военный институт иностранных языков, в 1956 г. – Военно-инженерную академию им. В.В. Куйбышева и до 1980 г. работал в ЦНИИ 27 Министерства обороны СССР, в 1980–2001 гг. – во Всероссийском институте научной и технической информации АН СССР, с 2003 г. – в лингвистической фирме МетаФраз. Им создана уникальная система машинного англо-русского и русско-английского фразеологического перевода (RETRANS), разработаны принципы морфологического и семантико-синтаксического анализа и синтеза русских текстов, построения языковых и программных средств информационных систем, в том числе систем машинного перевода с естественных языков. Под его руководством составлены многомиллионные машинные словари для автоматической обработки текстов, в частности систем автоматического индексирования документов. Им создана научная школа, насчитывающая десятки докторов и кандидатов наук.

Г.Г. Белоногов имеет более 100 научных трудов, среди которых монографии: «Автоматизированные информационно-поисковые системы» (1968, соавт. Р.Г. Котов); «Автоматизированные информационные системы» (1973, соавт. В.И. Богатырев); «Автоматизация процессов накопления, поиска и обобщения информации» (1979, соавт. А.П. Новоселов); «Языковые средства автоматизированных информационных систем» (1983, соавт. Б.А. Кузнецов); «Автоматизированная обработка научно-технической информации: Лингвистический аспект» (1984, соавт. Б.А. Кузнецов, А.П. Новоселов); «Компьютерная лингвистика и перспективные информационные технологии» (2004, соавт. Ю.П. Калинин, А.А. Хорошилов); «Семантические проблемы информатики» (2008) и др.

Это был целеустремленный человек, до последнего дня думавший о научных проблемах, заботившийся о сохранении и приумножении созданного им дела. Его отличала редкая порядочность, интеллигентность и внимательная заботливость в отношениях с работавшими с ним людьми.

Благодарная память о Герольде Георгиевиче сохранится у его многочисленных учеников и последователей.