МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

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. № 7 Москва 2019

ИНФОРМАЦИОННЫЙ ПОИСК

УДК 519.876.5:004.65

Е.М. Бениаминов

Алгебра скрытых отношений как средство моделирования статистических баз данных

Рассматривается аналог реляционной алгебры, используемый для моделирования баз данных в случае, когда пользователю значения атрибутов видны не полностью, а только с точностью до некоторого отношения эквивалентности на доменах. Этот случай важен для так называемых статистических баз данных, когда нужно выдавать и всесторонне анализировать сводную информацию по данным, хранящимся в некоторой базе данных, но доступ пользователям к персональным данным в базе данных закрыт. По аналогии с реляционным подходом в этом случае вводятся понятие скрытого отношения и система операций над скрытыми отношениями. Показаны некоторые соотношения между этими операциями и язык запросов.

Ключевые слова: *реляционная база данных, статистическая база данных, скрытые домены, операции над отношениями, группа симметрий, k-анонимность*

МОТИВАЦИЯ И НЕФОРМАЛЬНОЕ ВВЕДЕНИЕ

Проблема, которой посвящена настоящая работа не нова. В литературе давно обсуждается вопрос, как организовать доступ пользователям к базам данных для решения задач статистического анализа, не раскрывая при этом персональных данных, которые могут храниться в этих БД [1]. Эта тема не устарела и сейчас. Международные конференции "*Privacy in Statistical Databases*" («Конфиденциальность в статистических базах данных») проводятся уже в течение многих лет [2].

Один из подходов к статистическим базам данных состоит в том, что обычную реляционную БД, которая описывает классификацию объектов предметной

области, их свойства и связи между объектами, дополняют числовыми признаками-показателями, заданными на строках некоторых отношений. Запрос к такой базе данных состоит в построении таблиц агрегированных и статистических показателей на основе хранящихся в ней показателей и отношений. Такие статистические базы данных могут быть использованы для научных целей или для онлайн аналитических приложений (*OLAP)*. В настоящей работе предлагается развитие именно такого подхода.

Так же, как и в реляционном подходе к базам данных [3, 4], в этом подходе важными являются понятия отношения и атрибута отношения. Атрибут отношения задается именем *a* атрибута и областью значений (доменом) этого атрибута D_a. В отличие от стандартного подхода здесь будет предполагаться, что с одним и тем же именем атрибута могут связываться разные домены. При этом значения некоторых доменов могут наблюдаться полностью и применяться пользователями статистической базы данных в запросах. Такие домены D_a иногда будут называться наблюдаемыми. Предполагается также, что с именем атрибута *a* могут быть связаны и другие домены D'_a. Значения атрибутов в этих доменах могут не полностью наблюдаться пользователем в отношениях. Соответствие между ненаблюдаемыми и наблюдаемыми значениями задается отображением $p': D'_a \to D_a$. При этом предполагается, что каждое значение $d' \in D'_a$ пользователь воспринимает как значение $p'(d') \in D_a$ в области наблюдаемых значений D_a атрибута *a*.

Для примера в качестве атрибута рассмотрим атрибут с именем возраст *age* с наблюдаемыми значениями D_{age} в виде множества интервалов, которые именуются, как "младенческий", "младший дошкольный" и т. д. В качестве домена реальных значений для атрибута "возраст" может рассматриваться целочисленный тип данных $D'_{age} = \text{int}$. Отображение $p': D'_{age} \to D_{age}$ соответствует принадлежности числа к соответствующему интервалу.

Отношением *r* с набором атрибутов $\{a_1,...,a_n\}$ называется конечное подмножество $r \subset D'_{a_1} \times ... \times D'_{a_n}$ в декартовом произведении доменов соответствующих атрибутов. В нашем случае предполагается, что для статистических исследований строки отношений могут рассматриваться как в расширенных (скрытых) доменах D'_{a_i}, так и в наблюдаемых доменах. Отношения, в которых используются скрытые домены, естественно назвать скрытыми отношениями. Статистические показатели такие, как доход, количество штук, численность населения и т. д. задают веса на строках таких отношений.

Соединяя эти понятия, мы приходим к понятию показателя на скрытых отношениях, который представляется функцией (показателем), заданной на множестве строк отношения, где есть атрибуты с расширенными (скрытыми) доменами, и со значениями функции в множестве действительных чисел. Носителем такого показателя является множество строк, на которых этот показатель имеет ненулевое значение.

В настоящей работе мы попытаемся проанализировать ситуацию, в которой выполняется следующее основное предположение: пользователь знает имена отношений из схемы базы данных, может обратиться к скрытому отношению по имени, но видеть он может только суммарные значения весов строк по скрытым доменам самого отношения для строк наблюдаемой части носителя скрытого отношения. То же относится к скрытым отношениям, полученным в результате операций над ними, т. е. ответы на запросы пользователей к базе данных со скрытыми отношениями должны быть отношениями с наблюдаемыми доменами.

Чтобы прояснить эту ситуацию следует ввести систему операций над скрытыми отношениями и формально точно определить, что может наблюдать пользователь. Это и есть цель настоящей статьи.

ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

Введем основные определения понятий, которые нам потребуются для моделирования баз данных со скрытыми отношениями: основное – это понятие скрытого отношения, которое является обобщением обычного понятия отношения.

Так же, как и в реляционном подходе к базам данных, атрибут отношения — это пара *a* и D_a, где *a* – имя атрибута, D_a – область значений (домен) этого атрибута.

Набором атрибутов $A=\{(a_1, D_{a_1}),\ldots,(a_n, D_{a_n})\}$ называется последовательность атрибутов с различными именами.

Отношением *r* с набором атрибутов $A=\{(a_1, D_{a_1}),\ldots,(a_n, D_{a_n})\}$ называется произвольное подмножество в декартовом произведении доменов атрибутов, то есть $r \subseteq D_{a_1} \times ... \times D_{a_n}$.

Для приложений баз данных к статистическим исследованиям или для задач аналитического анализа понятие отношения нужно немного модернизировать.

<u>Определение 1.</u> *Взвешенным отношением (показателем) φ с набором атрибутов* $A=\{(a_1, D_{a_1}), \ldots, (a_n, D_{a_n})\}$ *называется произвольная функция* $\phi: D_{a_1} \times ... \times D_{a_n} \to R$, *где R – множество действительных чисел (с указанием единицы измерения). Через V(A) мы будем обозначать множество всех взвешенных отношений с набором атрибутов A, т.е.* $V(A) \stackrel{def}{=} \{\phi \mid \phi: D_{a_1} \times ... \times D_{a_n} \to R\}$. *Если* $(d_1,...,d_n) \in D_{a_1} \times ... \times D_{a_n}$ *– произвольная строка из декартового произведения доменов атрибутов, то значение функции* $\phi(d_1,...,d_n)$ *на этой строке называется весом строки* $(d_1,...,d_n)$ *в отношении φ.*

Множество $r_\phi \overset{def}{=} \{(d_1,...,d_n) : \phi(d_1,...,d_n) \neq 0\}$ *строк с ненулевым весом называется носителем взвешенного отношения* φ.

Нас, как правило, будут интересовать взвешенные отношения с конечным носителем.

Далее, по аналогии с реляционной алгеброй, мы должны определить операции над взвешенными отношениями.

Аналогами теоретико-множественных операций над отношениями являются операции сложения и умножения взвешенных отношений.

Определение 2. *Если* $\phi_1, \phi_2 : D_{a_1} \times ... \times D_{a_n} \to R$ *– два взвешенных отношения с одинаковыми наборами атрибутов* $A=\{(a_1, D_{a_1}),...,(a_n, D_{a_n})\}$, *то их суммой* $\phi_1 + \phi_2$, *разностью* $\phi_1 - \phi_2$ *и произведением* $\phi_1 \cdot \phi_2$ *называются взвешенные отношения с тем же набором атрибутов, веса которых на строках* $(d_1,...,d_n) \in D_{a_1} \times ... \times D_{a_n}$ *задаются, соответственно, выражениями:*

$$(\phi_1 + \phi_2)(d_1,...,d_n) \overset{def}{=} \phi_1(d_1,...,d_n) + \phi_2(d_1,...,d_n),$$
$$(\phi_1 - \phi_2)(d_1,...,d_n) \overset{def}{=} \phi_1(d_1,...,d_n) - \phi_2(d_1,...,d_n) \text{ и}$$
$$(\phi_1 \cdot \phi_2)(d_1,...,d_n) \overset{def}{=} \phi_1(d_1,...,d_n)\phi_2(d_1,...,d_n),$$

т. е. $\phi_1 + \phi_2$, $\phi_1 - \phi_2$ *и* $\phi_1 \cdot \phi_2$ *как функции являются суммой, разностью и произведением функций* φ_1 *и* φ_2.

Перейдем к определениям для построения операций, связывающих взвешенные отношения с разными наборами атрибутов.

Определение 3. *Пусть* $A=\{(a_1, D_{a_1}),...,(a_n, D_{a_n})\}$ *и* $B=\{(b_1, D_{b_1}),...,(b_k, D_{b_k})\}$ *– два набора атрибутов. Морфизмом* $\gamma : A \to B$ *из набора атрибутов A в набор атрибутов B называется отображение множеств имен атрибутов* $\gamma : \{a_1,...a_n\} \to \{b_1,...,b_k\}$, *которое мы будем обозначать той же буквой* γ, *вместе с отображениями доменов* $\gamma_{a_i} : D_{\gamma(a_i)} \to D_{a_i}$, *для каждого* $a_i \in \{a_1,...,a_n\}$.

Заметим, что, если $\gamma : A \to B$ морфизм наборов атрибутов, то по нему естественным образом строится отображение декартовых произведений

$$\Gamma_\gamma : D_{b_1} \times ... \times D_{b_k} \to D_{a_1} \times ... \times D_{a_n} \qquad (1)$$

по правилу: если $(h_{b_1},...,h_{b_k}) \in D_{b_1} \times ... \times D_{b_k}$, то

$$\Gamma_\gamma(h_{b_1},...,h_{b_k}) \overset{def}{=} (d_{a_1},...,d_{a_n}) \in D_{a_1} \times ... \times D_{a_n}, \qquad (2)$$

где $d_{a_i} = \gamma_{a_i}(h_{\gamma(a_i)})$

для $a_i \in \{a_1,...,a_n\}$.

Определение 4. *Если* $\phi : D_{a_1} \times ... \times D_{a_n} \to R$ *– произвольное взвешенное отношение с набором атрибутов* $A=\{(a_1, D_{a_1}),...,(a_n, D_{a_n})\}$, *и* $\gamma : A \to B$ *– морфизм атрибутов, где* $B=\{(b_1, D_{b_1}),...,(b_k, D_{b_k})\}$, *то через* $\gamma_*(\phi)$ *обозначим взвешенное отношение с набором атрибутов B, которое как функция* $\gamma_*(\phi) : D_{b_1} \times ... \times D_{b_k} \to R$ *является композицией функции* Γ_γ, *заданой выражениями (1) и (2), и функции* φ; *т. е.* $\gamma_*(\phi) \overset{def}{=} \phi \circ \Gamma_\gamma$, *где* $\circ$ *– знак композиции функций. Взвешенное отношение* $\gamma_*(\phi)$ *называется прямым образом взвешенного отношения* φ *относительно морфизма* γ.

Определение 5. *Если* $\psi : D_{b_1} \times ... \times D_{b_k} \to R$ *взвешенное отношение с набором атрибутов B, и* $\gamma : A \to B$ *– морфизм атрибутов, то через* $\gamma^*(\psi)$ *обозначим взвешенное отношение с набором атрибутов A, которое как функция* $\gamma^*(\psi) : D_{a_1} \times ... \times D_{a_n} \to R$ *на любой строке* $(d_{a_1},...,d_{a_n}) \in D_{a_1} \times ... \times D_{a_n}$ *задается выражением:*

$$\gamma^*(\psi)(d_{a_1},...,d_{a_n}) \overset{def}{=}$$
$$\overset{def}{=} 0 + \sum_{\substack{(h_{b_1},...,h_{b_k}) \in D_{b_1} \times ... \times D_{b_k} \\ where \quad \Gamma_\gamma(h_{b_1},...,h_{b_k}) = (d_{a_1},...,d_{a_n})}} \psi(h_{b_1},...,h_{b_k}). \qquad (3)$$

Взвешенное отношение $\gamma^*(\psi)$ *называется обратным образом взвешенного отношения* φ *относительно морфизма* γ.

Особенностью операции прямого образа γ_* в случае, когда морфизм атрибутов $\gamma : A \to B$ не сюръективен и домены D_{b_i} бесконечны, является то, что взвешенные отношения $\phi : D_{a_1} \times ... \times D_{a_n} \to R$ с конечным носителем могут переводиться этой операцией в скрытые отношения с бесконечным носителем. Для сглаживания этого недостатка операция прямого образа будет применяться в сочетании с операцией умножения на взвешенное отношение с конечным носителем, т. е. в виде $\psi\gamma_*(\phi)$, где $\psi : D_{b_1} \times ... \times D_{b_k} \to R$ – взвешенное отношение с конечным носителем.

Следующие свойства операций прямого и обратного образа, перечисленные в Утверждении 1, легко проверяются.

Утверждение 1. *Операции прямого и обратного образа над взвешенными отношениями удовлетворяют следующим соотношениям.*

1. Операции прямого и обратного образа функториальны. Это значит, что:

если $1_A : A \to A$ *— тождественное согласование атрибутов, то операции* $(1_A)^*$ *и* $(1_A)_*$ *являются тождественными отображениями;*

если $\gamma_1 : A \to B$, $\gamma_2 : B \to C$ *— морфизмы наборов атрибутов и* $\gamma_2 \circ \gamma_1 : A \to C$ *— композиция этих морфизмов, то выполняются следующие равенства* $(\gamma_2 \circ \gamma_1)_* = (\gamma_2)_* \circ (\gamma_1)_*$ *и* $(\gamma_2 \circ \gamma_1)^* = \gamma_1^* \circ \gamma_2^*$.

2. Для любого морфизма наборов атрибутов $\gamma : A \to B$ *операция прямого образа* $\gamma_* : V(A) \to V(B)$ *является гомоморфизмом алгебр взвешенных отношений относительно операций сложения и умножения. То есть* $\gamma_*(\phi_1 + \phi_2) = \gamma_*(\phi_1) + \gamma_*(\phi_2)$ *и* $\gamma_*(\phi_1\phi_2) = \gamma_*(\phi_1)\gamma_*(\phi_2)$, *для любых взвешенных отношений* $\phi_1, \phi_2 \in V(A)$ *с набором атрибутов* $A = \{a_1, ..., a_n\}$.

3. Для любого морфизма наборов атрибутов $\gamma : A \to B$ *операция прямого образа* $\gamma^* : V(B) \to V(A)$ *перестановочна с операцией суммы, т. е.* $\gamma^*(\psi_1 + \psi_2) = \gamma^*(\psi_1) + \gamma^*(\psi_2)$ *для любых взвешенных отношений* $\psi_1, \psi_2 \in V(B)$ *с набором атрибутов B, но не сохраняет произведения, т. е., в общем случае* $\gamma^*(\psi_1\psi_2) \neq \gamma^*(\psi_1)\gamma^*(\psi_2)$.

Утверждение 2. *Для любого морфизма наборов атрибутов* $\gamma : A \to B$ *и любых взвешенных отношений* $\psi \in V(B)$ *и* $\phi \in V(A)$ *с конечными носителями и с наборами атрибутов* $B = \{b_1, ..., b_k\}$ *и* $A = \{a_1, ..., a_n\}$, *соответственно, выполняется следующее равенство:* $\gamma^*(\gamma_*(\psi)\phi) = \psi\gamma^*(\phi)$.

Доказательство Утверждения 2 проводится непосредственной проверкой равенства с использованием определений операций. Пусть $d_B : \{b_1, ..., b_k\} \to D_{b_1} \cup ... \cup D_{b_k}$ – произвольная строка из декартового произведения $D_{b_1} \times ... \times D_{b_k}$, тогда по Определениям 4 и 5 операций прямого и обратного образа взвешенного отношения, а также из свойства дистрибутивности умножения относительно сложения, из левой части равенства Утверждения 2 на этой строке получаем:

$$
\begin{aligned}
(\gamma^*(\gamma_*(\psi))\phi)(d_B) &\overset{def}{=} 0 + \sum_{\substack{d_A \in D_{a_1} \times ... \times D_{a_n} \\ where \Gamma_\gamma(d_A) = d_B}} (\gamma_*(\psi)\phi)(d_A) = \\
&= 0 + \sum_{\substack{d_A \in D_{a_1} \times ... \times D_{a_n} \\ where \Gamma_\gamma(d_A) = d_B}} \psi(\Gamma(d_A))\phi(d_A) = \\
&= \psi(d_B)\left(0 + \sum_{\substack{d_A \in D_{a_1} \times ... \times D_{a_n} \\ where \Gamma_\gamma(d_A) = d_B}} \phi(d_A)\right) = \\
&= \psi(d_B)(\gamma^*(\phi)(d_B)).
\end{aligned}
\tag{4}
$$

В результате получено значение правой части требуемого равенства Утверждения 2 для строки $d_B \in D_{b_1} \times ... \times D_{b_k}$.

Определение 6. *Если* $r \subseteq D_{a_1} \times ... \times D_{a_n}$ *– отношение, в котором один из атрибутов (например,* a_n*) функционально зависит от остальных атрибутов* $\{a_1, ..., a_{n-1}\}$ *и домен* D_{a_n} *этого атрибута является числовым, то через* $\psi(r, a_n)$ *обозначается показатель* $\psi(r, a_n) : pr(r \mid a_1, ..., a_{n-1}) \to D_{a_n}$, *заданный на проекции* $pr(r \mid a_1, ..., a_{n-1})$ *отношения r на множество атрибутов* $\{a_1, ..., a_{n-1}\}$, *который каждой строке* $(d_1, ..., d_{n-1})$ *из проекции* $pr(r \mid a_1, ..., a_{n-1})$ *выдает значение* $\psi(r, a_n)(d_1, ..., d_{n-1}) = d_n$ *строки* $(d_1, ..., d_{n-1}, d_n)$ *отношения r.*

Определим операцию, обратную только что определенной.

Определение 7. *Если* $\psi : D_{a_1} \times ... \times D_{a_{n-1}} \to R$ *– взвешенное отношение с носителем* $r \subseteq D_{a_1} \times ... \times D_{a_{n-1}}$, и a_n *– имя атрибута, которого нет среди имен атрибутов отношения r, то через* (r, a_n, ψ) *обозначается отношение* $(r, a_n, \psi) \subset D_{a_1} \times ... \times D_{a_{n-1}} \times R$, *состоящее из строк вида* $(d_1, ..., d_{n-1}, \psi(d_1, ..., d_{n-1}))$ *для всех строк* $(d_1 ..., d_{n-1}) \in r$.

Следующая операция важна для составления по некоторому отношению *r* сводных отчетов-отношений с использованием стандартного набора операторов, принимающих числовые значения, и операторов над числами.

Обычно, как, например, в языке запросов SQL, используются следующие стандартные операторы:

count(| [distinct] имя_атрибута)* – вычисления числа строк или числа различных значений в выделенном атрибуте;

max(имя_атрибута), min(имя_атрибута), sum([distinct] имя_атрибута), avg([distinct] имя_атрибута) – вычисления, соответственно, минимального, максимального, суммарного или среднего значения по некоторому атрибуту, принимающему, числовое значение (для всех или только различных значений этого атрибута в зависимости от присутствия ключевого слова *distinct* в соответствующем операторе).

Для применения операции *GROUP BY* к отношению *r* с набором атрибутов *A* нужно указать подмножество атрибутов $B \subseteq A$, по которому следует произвести группирование (разбиение множества строк исходного отношения на подмножества с равными значениями в подстроках с атрибутами *B*), и некоторый оператор, из перечисленных выше, с указанием атрибута из *A*, не входящего в *B*. Результатом такой операции является взвешенное отношение-показатель с набором атрибутов *B*, функция весов строк которого вычисляется указанным оператором по атрибуту в каждой группе.

Помимо приведенных выше операций допускаются также все операции Кодда над отношениями, включая теоретико-множественные операции и операции проекции, соединения и переименования атрибутов в отношениях.

Рассмотрим далее базу данных с отношениями $r_1, ..., r_n$, с именами этих отношений $R_1, ..., R_n$ и доменами атрибутов отношений $D_1, ..., D_k, D'_1, ..., D'_s$. Домены $D_1, ..., D_k$ называются наблюдаемыми, а до-

мены $D'_1,...,D'_s$ – скрытыми. К наблюдаемым относятся также некоторые числовые домены.

Определение 8. *Скрытым отношением r называется отношение, содержащее атрибуты со скрытыми доменами. Запросом к базе данных называется правильно построенное выражение из $R_1,...,R_n$ и имен операций над отношениями, задающее композицию из определенных выше операций с этими отношениями. Ответом на запрос к базе данных в текущем состоянии $r_1,...,r_n$ называется результат вычисления выражения запроса после подстановки в запрос вместо имен отношений $R_1,...,R_n$ их состояний $r_1,...,r_n$..*

Определение 9. *Запрос к базе данных со скрытыми отношениями называется внешним, если ответом является отношение с наблюдаемыми доменами атрибутов.*

СИММЕТРИЯ СКРЫТЫХ ОТНОШЕНИЙ

Пусть, как и ранее, база данных задается отношениями $r_1,...,r_n$ с именами этих отношений $R_1,...,R_n$ и доменами атрибутов отношений $D_{t_1},...,D_{t_k}, D'_{s_1},...,D'_{s_m}$, где $T=\{t_1,...,t_k\}$ – множество всех типов наблюдаемых доменов базы данных, и $S=\{s_1,...,s_m\}$ – множество всех типов скрытых доменов базы данных. Обозначим через $D=D_1\cup...\cup D_k$ разъединенное объединение всех наблюдаемых доменов, через $D'=D'_1\cup...\cup D'_s$ – разъединенное объединение всех скрытых доменов базы данных, а через $D^{all}=D\cup D'$ – объединение всех доменов. Естественно, что каждый элемент $d\in D^{all}$ имеет свой тип $type(d)\in T\cup S$, заданный отображением $type: D^{all}\to T\cup S$, которое определяется равенством $type(d)=t_j$, если $d\in D_{t_j}$, для $j=1,...,k$ или $type(d)=s_i$, если $d\in D'_{s_i}$, для $i=1,...m$. Через D_c^{all} обозначим подмножество в D^{all}, состоящее из всех элементов, входящих в строки отношений $r_1,...,r_n$. Множество D_c^{all} естественно назвать доменом – носителем базы данных $r_1,...,r_n$.

Предположим также, что среди отношений $r_1,...,r_n$ схемы базы данных нет отношений порядков на скрытых доменах.

Определение 10. *Биективное отображение $h: D_c^{all}\to D_c^{all}$, оставляющее на месте элементы наблюдаемых доменов и переставляющее элементы скрытых доменов, но сохраняющее типы элементов, называется симметрией базы данных $r_1,...,r_n$, если для любой строки $(d_1,\ldots,d_q)\in r_i$ любого отношения $r_i\in r_1,\ldots,r_n$ строка $(h(d_1),\ldots,h(d_q))$ также принадлежит этому же отношению r_i, т. е. $(h(d_1),\ldots,h(d_q))\in r_i$.*

Так как отображение симметрии *h* – биективно, то оно, по определению, и инъективно, и, следовательно, различные строки отношения r_i переводятся в различные строки этого же отношения. Так как отношение r_i конечно, то из инъективности *h* на строках следует, что *h* биективно действует на строках каждого отношения $r_i\in r_1,\ldots,r_n$.

Легко проверить, что каждая симметрия базы данных есть и симметрия отношения, которое является ответом на запрос к этой базе данных (предполагается, что в запросах не используются константы из скрытых доменов и отношения, отличные от $r_1,\ldots,r_n$, содержащие атрибуты со скрытыми доменами). С этой целью нужно проконтролировать выполнение этого свойства для всех перечисленных операций над отношениями (так как каждый запрос – это композиция элементарных операций над отношениями).

Нетрудно доказать также, что,

если $h: D_c^{all}\to D_c^{all}$ является симметрией базы данных, то и обратное отображение $h^{-1}: D_c^{all}\to D_c^{all}$ – это симметрия;

если $h_1,h_2: D_c^{all}\to D_c^{all}$ – две симметрии базы данных, то композиция этих отображений $h_1\circ h_2: D_c^{all}\to D_c^{all}$ также является симметрией базы данных;

тождественное отображение $1_{D_c^{all}}: D_c^{all}\to D_c^{all}$, очевидно, тоже симметрия базы данных.

Определение 11. *Множество всех симметрий состояния r_1, ..., r_n базы данных вместе с этими операциями над симметриями образуют группу, которая называется группой симметрий базы данных (в данном состоянии).*

Обозначим эту группу через *H*.

Определение 12. *Если H – группа симметрии базы данных со скрытыми отношениями и $d'\in D'_j$ – элемент скрытого домена, то множество $H(d')\overset{def}{=}\{d''\in D'_j \mid d''=h(d') \text{ для } h\in H\}$ называется траекторией элемента d' относительно группы H. Элемент $d'\in D'_j$ называется неподвижным, если множество H(d') состоит из одного элемента.*

Если группа *H* симметрий базы данных со скрытыми отношениями оставляет неподвижными элементы наблюдаемых доменов (т. е. $h(d)=d$ для любых $d\in D$ и любого $h\in H$), то она оставляет неподвижными и строки отношений, которые являются ответами на внешние запросы к базе данных. Это следует из того, что такие отношения по определению не содержат атрибутов со скрытыми доменами.

Пусть *r* – отношение, которое является ответом на некоторый запрос к состоянию базы данных $r_1,...,r_n$. Допустим также, что отношение *r* имеет атрибуты со значениями в скрытых доменах. Если группа *H* – группа симметрий этой базы данных, то все элементы группы *H,* как было отмечено, являются симметриями отношения *r*, т. е. группа *H* действует на строках отношения *r*. Более того, легко видеть: если группа *H* действует на элементы скрытых доменов без неподвижных элементов, то и на строках отношения *r* она действует без неподвижных элементов. Отсюда следует важное для баз данных утверждение.

Утверждение 3. *Если у группы H симметрий базы данных нет неподвижных элементов в скрытых доменах, то по внешним запросам невозможно однозначно определить строку отношения со скрытыми доменами, являющегося ответом на запрос к этой базе данных, т. е.: если отношение r – ответ на запрос к базе данных и строка* $\bar{d}$ *принадлежит r, то множество строк* $H(\bar{d}) = \{h(\bar{d}) \mid h \in H\}$, *полученных действием симметрий из группы H на строку* $\bar{d}$, *является подмножеством в r, т.е.* $H(\bar{d}) \subseteq r$.

Таким образом, если отношение *r* является ответом на запрос к базе данных и содержит атрибуты со скрытыми доменами, то число строк в $H(\bar{d})$ для любой строки $\bar{d} \in r$ не меньше, чем число $k = \min_{d' \in D'} |H(d')|$ – минимальное число элементов в траекториях элементов скрытых доменов. Отсюда следует, что при построении баз данных со скрытыми отношениями для увеличения степени неопределенности распознавания скрытых элементов отношений необходимо увеличивать число *k,* которое в публикациях по безопасности баз данных и социальных сетей получило название *k*-анонимность (*k-anonymity*) [5, 6].

Теперь рассмотрим, в какой степени множества элементов наблюдаемых доменов определяют множества элементов скрытых отношений в базе данных со скрытыми доменами атрибутов.

Пусть, как и ранее, $r_1,...,r_n$ – состояние базы данных со скрытыми отношениями и $M \subseteq D$ – подмножество во множестве *D* элементов наблюдаемых доменов. Заменим все элементы наблюдаемых доменов, не входящих в подмножество *M*, в строках отношений $r_1,...,r_n$ на новый элемент *NULL.* Получим отношения $r_1^M,...,r_n^M$, имеющие атрибуты, значения которых лежат во множестве $D^M = M \cup \{NULL\} \cup D'$, где *D'*, как и ранее, – объединение всех скрытых доменов.

Пусть H^M – группа симметрии состояния $r_1^M,...,r_n^M$ базы данных с доменом D^M и скрытым доменом *D'* (см. Определение 11), т. е. H^M – множество всех биекций $h: D^M \to D^M$, сохраняющих типы элементов и оставляющих элементы множества $D^M \cup \{NULL\}$ на месте, которые переводят отношения $r_1^M,...,r_n^M$ в себя. В частном случае, когда *M=D,* группа $H^D = H$.

Соответственно, $H^M(d')$, где $d' \in D' \subset D^M$, – это траектория элемента *d'* скрытого домена относительно группы H^M, действующей на множестве D^M. Аналогично $k = \min_{d' \in D'} |H(d')|$ – минимальному числу элементов в траекториях элементов скрытых доменов относительно группы *H*, определим $k^M = \min_{d' \in D'} |H^M(d')|$ – минимальное число элементов в траекториях элементов скрытых доменов относительно группы H^M.

Число k^M – это минимальный размер подмножества в скрытых доменах, определяемого подмножеством наблюдаемых элементов $M \subseteq D$ в состоянии базы данных $r_1,...,r_n$.

БЛИЗКИЕ РАБОТЫ

Проблема, как организовать доступ пользователям к базам данных для задач статистического анализа, не раскрывая персональных сведений, которые могут храниться в этих базах данных, актуальна и сегодня. Первая статья по *k*-анонимности появилась в 1998 г. [5], более подробно этот подход изложен в [6]. Затем поток публикаций на тему *k*-анонимности стал значительно расти. Связь *k*-анонимности с симметрией данных в социальных сетях обсуждалась в [7]. Современное состояние исследований, связанных с анонимностью в базах данных и сетях, *k*-анонимностью, а также с некоторыми проблемами конфиденциальности, возникающими в результате анализа Больших Данных, показано в [8].

ЗАКЛЮЧЕНИЕ

Объектом исследования в настоящей работе являются реляционные базы данных, в которых некоторые атрибуты принимают значения в скрытых доменах. Предполагается, что пользователи таких баз данных могут обращаться к ним с произвольными запросами, но значения атрибутов в скрытых доменах не выдаются в качестве ответов на запросы к таким базам данных.

Нами рассмотрены взвешенные отношения в реляционных базах данных (показатели) и операции над такими отношениями, а также приведены некоторые соотношения между этими операциями.

Основная проблема, которая возникает при работе с базами данных, содержащими скрытые отношения, – это как обеспечить невозможность вывода значений атрибутов в скрытых доменах этой базы по любым допустимым запросам к ней. Мы предлагаем подход к этой проблеме, основанный на применении понятия группы симметрий состояния базы данных. Этот подход позволил найти условия, при которых элементы некоторых множеств в скрытых доменах невозможно различить с помощью внешних запросов к этой базе данных.

СПИСОК ЛИТЕРАТУРЫ

1. Jaideep Srivastava, Hung Q. Ngo. Statistical databases // In Wiley Encyclopedia of Electrical and Electronics Engineering. – Hoboken, NY: Wiley, 1999. – URL: https://onlinelibrary.wiley.com/doi/book/10.1002/047134608X
2. Privacy in Statistical Databases “UNESCO Chair in Data Privacy, International Conference, PSD 2018, Valencia, Spain, September 26–28, 2018” // Proceedings, Springer, LNCS, 2018. – Vol. 11126. – URL: https://www.springer.com/gp/book/9783319997704

3. Мейер Д. Теория реляционных баз данных. – М. : Мир, 1987. – 608 с.
4. Бениаминов Е.М. Алгебраические методы в теории баз данных и представлении знаний. – М.: Научный мир, 2003. – 184 с.
5. Samarati P., Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression // Technical Report SRI-CSL-98-04, Computer Science Laboratory, SRI International, 1998. – URL: http://www.csl.sri.com/papers/sritr-98-04/
6. Sweeney L. *K*-anonymity: a model for protecting privacy // International Journal on Uncertainty, Fuzziness and Knowledge-based Systems. – 2002. – Vol. 10(5). – P. 557-570.
7. Wu W., Xiao Y., Wang W., He Z., Wang Z. *K*-symmetry model for identity anonymization in social networks // Advances in Database Technology – EDBT 2010 Proceedings of the 13th International Conference on Extending Database Technology Lausanne, Switzerland, March 22–26, 2010. – P. 111–122.
8. Salas J., Domingo-Ferrer J. Some Basics on Privacy Techniques, Anonymization and their Big Data Challenges // Math. Comput. Sci. – 2018, – Vol.12(3). – P. 263–274.

Материал поступил в редакцию 22.04.19

Сведения об авторе

БЕНИАМИНОВ Евгений Михайлович – доктор физико-математических наук, профессор, заведующий кафедрой Российского государственного гуманитарного университета (РГГУ), Москва.
e-mail: ebeniamin@yandex.ru

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 004.91:81’37

Н.В. Максимов, О.Л. Голицына, К.В. Монанков, А.А. Лебедев, Н.А. Баль, С.Г. Кюрчева

Средства семантического поиска, основанные на онтологических представлениях документальной информации

Рассматриваются онтологические средства идентификации и представления семантики текстовых документов применительно к задачам интерактивного информационного поиска. Представлены реализации операций над онтологиями, позволяющие формировать образы новых смыслов в предметной области. Для реализации операций, а также для идентификации нечетких соотношений (в том случае, когда сущности и связи заданы на разном уровне общности и/или выражены разными лингвистическими конструкциями) используются таксономии отношений (парадигматических и синтагматических) и сущностей (политематический тезаурус понятий).

Предложены интерактивные средства, использующие операцию построения аспектных проекций для графовых представлений онтологий, позволяющие снизить размерность графа до уровня, приемлемого с точки зрения возможностей отображения и восприятия. Рассмотрены возможности контекстно обусловленного использования сущностей и отношений для развития процесса поиска. Обсуждается использование графов онтологий организационных процессов в качестве «навигационной карты», обеспечивающей в графическом интерфейсе новые точки входа – объекты, которые будут представлять собой контекстно определенные паттерны поисковых запросов в задачах отбора по сходству, либо в задачах экспертного анализа, что расширяет систему механизмов поиска и позволяет не только увеличить его полноту и точность, но и сделать схему поисковой навигации более естественной, приблизив ее к схемам понимания и синтеза знаний.

Ключевые слова: *семантический поиск, интерактивный графический интерфейс, обработка текста, графовые представления онтологий, операции над онтологиями*

ВВЕДЕНИЕ

Семантический поиск в настоящее время ассоциируется с двумя классами задач. Первый – это процесс отбора документов, отвечающих информационной потребности, выраженной запросом. Это хорошо известные механизмы – фактографический и тематический поиски, поиск аналогов, которые в совокупности нацелены на снятие неопределенности объекта/предмета поиска (лингвистической, семантической, прагматической). Второй класс – это комплексный аналитический поиск, который ориентирован на задачи синтеза нового знания. Сюда же относятся задачи мониторинга проектов; выявление оснований, ограничений и потенциально возможных связей; анализ новизны; задачи оценки соответствия документации установленным критериям и нормативным требованиям и т.п. Решения задач второго класса основываются на соотнесении и систематизации информации и характеризуются комплексностью результата и вариантностью его использования. Соответственно, и развитие поисковых технологий идет по пути создания аналитических и синтетических инструментов, обеспечивающих, в том числе, поиск зависимостей и различий. Естественность этой тенденции в том, что организационные и технологические границы между основной и информационной деятельностью практически становятся незаметными. С одной стороны, за счет автоматизации всё больше функций исследовательской, конструкторской, проектной, управленческой и т.п. деятельности теперь выполняются в информационно-вычислительной среде. С другой стороны – все больше информационных ресурсов становятся оперативно доступными исполнителю. Интеграция этих двух составляющих даже без уровня данных и семантики, а только на уровне размещения ИТ-инструментов на одном компьютере,

уже изменяет стиль работы, увеличивая оперативность и детальность проработки, а также обоснованность оценки создаваемых решений. А возможность подключения баз знаний создает условия автоматизации процессов взаимодействия с информационными ресурсами по таким функциям, как формулирование запроса и систематизация выдачи. Отметим, что в рамках когнитивного процесса эти функции соответствуют принципиально важным этапам – постановке задачи и синтезу целевого результата.

Онтологический подход позволяет представить семантику описанного в документе отдельного решения системой понятий и связей, т.е. при поиске будут использоваться завершенные смысловые конструкции. В частности, граф будет представлять пространство точек входа в информационный массив, обеспечивая возможность перехода от вершины к фрагментам текста документа.

В настоящей статье рассматриваются некоторые результаты разработки и использования онтологических средств представления и идентификации смысла применительно к задачам поиска документальной информации. Такие средства синтетически соединяют лингвистические, статистические и когнитивные подходы, а технологической основой является графовое представление онтологий предметной области и онтологий процессов деятельности в качестве упорядоченного множества точек входа, обеспечивающих задание направленности поискового процесса и его контроль за счет явно формируемого и схематично предоставляемого пользователю контекста ситуации.

Настоящая работа основывается на результатах научных и практических исследований авторов в области информационного поиска, онтологического моделирования и человеко-машинных интерфейсов[1]. Экспериментальные исследования (в том числе в составе опытной и опытно-промышленной эксплуатации) проводились на массивах полных текстов научных (статьи, диссертации, научные отчеты), технических (проектная и конструкторская документации), эксплуатационных (отчеты и расследования о неисправностях и т.д.), нормативных (стандарты, регламенты) и других (например, судебные решения) документов на платформе информационно-аналитической системы *xIRBIS*[2].

СЕМАНТИКА В ПОИСКОВЫХ ПРОЦЕССАХ

Основой поиска являются технологии индексирования, где индексы – это точки входа для поиска. Традиционно (и вполне обоснованно) в качестве поисковых индексов используются имена понятий (или объектов, свойств и т.д.), выделяемых из текста. В случае же использования онтологий индексами могут быть и связи (точнее, их имена). В отличие от линейного поискового образа онтология, представляющая глубинные семантические связи, отражает конкретность ситуации, т.е. мы переходим от точечного (теоретико-множественного) представления смысла в документах и запросах к сетевому.

В части технологий взаимодействия с поисковой системой также наблюдается тенденция перехода от традиционной вербальной формы запроса к графической. Например, в [4] для расширения запроса предлагается использовать семантические сети. Термины запроса сопоставляются с терминами семантической сети, из которой выбираются понятия для формирования дополнительных условий поиска. В [5] предлагается визуальное представление лексического окружения запроса, что позволяет пользователю интерактивно уточнять запрос, выбирая те термины, которые наиболее точно отражают информационную потребность. В [6] рассматривается метод, основанный на приведении временного ряда (как паттерна запроса) к дескриптивной форме, что позволяет использовать традиционные поисковые механизмы, а интерфейсные средства дают возможность пользователю не только анализировать динамику и статистические связи документальных потоков, но и, обращаясь к элементам и фрагментам временного ряда, манипулировать соответствующими выборками документов.

Особое внимание при разработке поисковых интерфейсов уделяется когнитивной нагрузке пользователя. Например, в [7] рассматриваются некоторые принципы визуализации информации, сокращающие когнитивную нагрузку при решении поисковых задач, и приводятся примеры современных интерактивных интерфейсов. С целью минимизации нагрузки на память пользователя в [8, 9] предлагается в каждый момент времени обеспечивать доступ только к тем возможностям интерфейса, которые полезны для решения текущей задачи. В этом направлении качество поиска может быть улучшено (как показано в [10, 11]) также и включением в поисковый процесс алгоритмов персонализации, основанных на анализе поведения пользователя.

Помимо трудностей при формировании запроса пользователь сталкивается с проблемой сложности оценки получаемых результатов, источниками которой обычно являются отсутствие средств развития запроса и тривиальность представления результатов поиска (чаще всего в виде статических списков). Одно из решений этой проблемы описано в [12], где предложен интерактивный инструмент, который при поиске отображает частотные распределения терминов найденных документов. Это позволяет пользователям оперативно сортировать результаты поиска, а также модифицировать выражение запроса для получения новых результатов.

Существенным инструментом поиска становится визуальная аналитика, в частности, использующая графические образы объектов, процессов и ситуаций в предметной области. Примерами такого рода решений являются когнитивные карты, а также системы визуализации больших семантических данных: *IBM i2* [13], например, предлагает большое количество удобных инструментов для формирования различных когнитивных схем взаимосвязей меж-

[1] Основные положения и обзоры литературы по онтологиям приведены в [1], по методологии построения и использования онтологий – в [2], по подходам к типизации отношений – в [3].

[2] Информационно-аналитическая система xIRBIS используется рядом ведущих информационных центров и организаций для создания промышленных документальных баз данных.

ду именованными объектами в виде графов; *Palantir* [14] обеспечивает визуализацию больших массивов данных из разнородных источников, что позволяет находить взаимосвязи между объектами, обнаруживать совпадения между объектами и событиями, выявлять аномальные объекты. Такого рода средствами располагают и системы типа *RCO Fact Extractor* [15], *ABBYY Compreno* [16], которые на основе лингвистического и семантического анализа текстов формируют графовые представления ситуаций. Отметим, что упомянутые системы (по крайней мере, в части наиболее известных классов задач, как например, анализ и предотвращение финансовых преступлений и терроризма, предиктаналитика в экономике, и т.п.) представляют отдельные ситуации, а не *процессы*[3].

Таким образом, тренды в направлении достижения более высоких показателей информирования (полнота, точность, эффективность) связываются с внедрением методов глубокого индексирования и предоставлением пользователю более широкого разнообразия точек входа, методов поиска и средств анализа результатов, с одной стороны, а с другой – с использованием средств снижения когнитивной нагрузки и сокращения размерности[4] операционного пространства, в том числе, путем разработки адаптивных интерфейсов, учитывающих характер задач, решаемых пользователем, и его поведение.

Далее рассмотрим методы и средства (в частности, использующие операцию аспектной проекции), которые позволяют конструктивно сократить размерность отображаемого графа автоматически получаемой онтологии, представляя объекты и связи предметной области на большем или меньшем уровне обобщенности или в некотором аспекте.

В контексте принятого в программной инженерии понятия «точка входа» можно отметить, что существо интерактивного семантического поиска состоит в формировании множеств таких точек входа и к тем данным (понятиям, документам или их фрагментам), которые в совокупности со специфицированной информационной потребностью пользователя будут играть роль информации [18], т.е. обретут свойство потенциальной *полезности* для решения целевой задачи пользователя. При этом отметим, что понятие информации может рассматриваться в двух аспектах:

1) как совокупность сведений (описание объекта поиска), которых не хватало пользователю для построения конкретного целевого результата его основной деятельности (информационный поиск, своими результатами замещающий основную деятельность);

2) как данные для принятия решения по выбору того или иного замещающего объекта.

В отличие от первого аспекта (который сводится к фактографическому поиску и потому нами не рассматривается), модели и технологии информационного обеспечения акта выбора замещающего объекта должны учитывать факторы, обуславливающие его неоднозначность, условность и стохастичность.

1. Выбор замещающего объекта, с одной стороны, имеет некоторую предысторию (представленную теми или иными фактами), а с другой – может иметь соответствующие последствия (разные и с разной временной и процессной отдаленностью), т.е. для того, чтобы сформировать семантически точный результат, система должна иметь и предоставлять пользователю и средства управляемого расширения запроса как в предметной, так и в понятийной сфере[5], и средства сужения и систематизации выдачи, как на уровне массива записей, так и на уровне отдельного документа через выделение фрагментов полного текста, семантически точно соответствующих смыслу запроса.

2. Необходимо учитывать семиотическую природу документальной формы информации. Это означает, что система должна иметь средства контролируемого последовательного снятия синтаксической, семантической и прагматической неопределенности предмета поиска, в итоге обеспечивая согласование лексики и контекста пользователя и системы.

3. Для информационной поддержки задач типа перспективных исследований (главным назначением которых является создание нового, т.е. неизвестного) в принципе невозможно априори задать критерий выбора, как и критерий останова процесса уточнения/расширения запроса. Здесь остается единственный вариант – ориентироваться по признакам информационного насыщения или исчерпания прироста новой информации/терминологии в очередных выдачах. Это означает, что система должна иметь средства накопления результатов поиска, а также выборочного их использования и сопоставления.

Система, непосредственно удовлетворяющая этим требованиям, будет предоставлять существенно больше «точек входа», чем человек способен конструктивно воспринять, т.е. она должна иметь и предоставлять пользователю также и возможность эти точки «фильтровать» и упорядочивать.

Отсюда следует, что онтология (как инструмент), помимо классического свойства «быть средством концептуализации», должна обладать свойствами, отражающими, в том числе, фактографию отдельных решений (на уровне деталей) и динамику использования как отдельных понятий, так и понятийных конструкций. Кроме того, в идеале онтология должна строиться по полным текстам полного комплекта документов, относящихся ко всем этапам жизненного цикла объекта. Такая онтология в пределе может представляться как подобие «сборочного чертежа», обеспечивающего построение *образа* описываемого объекта в когнитивном пространстве приемника (основные подходы и обзоры см. в [1, 2]).

[3] Сложность построения графовых представлений процессов синтеза нового знания обусловлена еще и тем, что здесь нет предопределенных типовых схем решения

[4] Условия взаимодействия пользователя с системой подчиняется также и принципу Миллера [17], отражающему количественные ограничения на способность пользователя к восприятию и идентификации информации.

[5] Отсюда следует, что система должна располагать полными и глубокими знаниями, представленными хорошо формализованными объектами. На сегодня – это тезаурусы, классификации, таксономии, онтологии предметной области.

ОНТОЛОГИЧЕСКИЕ ПРЕДСТАВЛЕНИЯ И ИХ ИСПОЛЬЗОВАНИЕ ПРИ ПОИСКЕ

Рассмотрим онтологический подход к решению таких проблем документального информационного поиска в больших массивах разнородной информации, как сокращение размерности пространства объектов, подлежащих анализу человеком, и увеличение оперативности и адекватности оценки получаемых результатов поиска. В основе решения этих проблем лежат глубокое многоаспектное индексировании текста в виде графовых образов (при соблюдении требований повышения полноты и точности отбора) и организация доступа с использованием навигации по графу (в том числе и соответствующие интерфейсные решения).

Онтологии: определение, операции и представление

Используемое в настоящей работе определение онтологии[6], отражающее динамизм и семиотическую природу познания, в [1] дано с позиций Общей теории систем как совокупность трёх взаимосвязанных систем[7].

Автоматическое построение онтологий для текстов на естественном языке, т.е. выявление отношений между выделяемыми в тексте понятиями (сущностями), основывается на преобразовании лингвистических отношений в функциональные (подробнее см. [19]). При этом набор типовых, так называемых функциональных, связей для предметной области (ПрО) может быть типизирован и обычно ограничен (подробнее см. [3]).

Представление онтологии на структурном уровне в виде графов позволяет формализовать операции над онтологиями на основе теоретико-графовых аксиом [1]. Основными операциями являются: бинарные – объединения, пересечения, проекции и унарные – масштабирования онтологий[8].

В качестве модели данных онтологии используется взвешенный ориентированный граф, обладающий в общем случае (вследствие семиотической и лингвистической природы онтологии) свойствами мета-, гипер- и мультиграфа[9].

Свойства метаграфа определяются наличием метавершин, каждая из которых представляет некоторый подграф (например, разрешение анафорических ссылок или раскрытие содержания путем использования определения). Для графовых форм, отображающих семантику текстов (и познания), наличие метавершин вполне конструктивно и естественно. Такая вершина, соответствующая (по своему имени) сущности – понятию, узлу, композиции и т.п., выступает в качестве атомарного семантического эквивалента смысла, определяемого неатомарной конструкцией (выражением).

Граф онтологии, построенной по тексту, представляющему нетривиальное описание какого-либо решения или гипотезы, должен обладать свойствами гиперграфа, поскольку имя отдельной сущности (объекта, понятия) может быть связано с некоторой другой сущностью отношениями нескольких типов, так как эта пара сущностей в разных частях текста может быть участником разных ситуаций, в которых сущности будут иметь разные роли.

Наличие свойств мультиграфа определяется тем, что онтология, являясь формой информации, обладает свойством эмерджентности и межотраслевым свойством, т.е. данные (и текст, как носитель информации), составляющие онтологию, интерпретируются в конкретном контексте, а именно целевой задачи и в контексте наличных знаний пользователя, формируя в итоге конкретное знание. Но, поскольку содержание текста может использоваться для решения более чем одной задачи, это означает, что по одному тексту будет, по существу, построено несколько онтологий (мультионтология), соответствующих разным решениям (контекстам, аспектам и т.п.). Таким образом, можно считать, что на общем (одном и том же) множестве сущностей и отношений, выделяемых из одного текста, будет построен мультиграф – множество графов, каждый из которых представляет свою целостную картину (онтологию).

Отметим, что ориентированность графа онтологии[10] определяется не только ориентированностью

---

[6] Онтология определяется как совокупность трёх взаимосвязанных систем: $O =< S_f, S_c, S_t \equiv >$.
S_f – функциональная система (объекты и связи действительности) определяется как $S_f =< M_f, A_f, R_f, Z_f >$, где M_f – множество объектов (сущностей), A_f – множество характеристических свойств, R_f – множество функциональных отношений, представленных типизированными ситуативными связями, характерными для ПрО, Z_f – закон композиции, т.е. правил и схем упорядочения объектов (например, таксономия ПрО).
S_c – понятийная система, определенная как $S_c =< M_c, A_c, R_c, Z_c >$, где M_c – множество понятий ПрО, A_c – множество признаков систематизации понятий (мерономия ПрО), R_c – множество отношений (прежде всего, парадигматических), определяющих классы/подклассы, Z_c – закон композиции (схема представления, определяющая какие понятия, в какой взаимосвязи, и в каком порядке будут включаться в классы).
S_t – терминологическая система, определенная как $S_t =< M_t, A_t, R_t, Z_x >$, где M_t – множество терминов, A_t – множество свойств, R_t – множество отношений эквивалентности и включения, а также лингвистических отношений, Z_t – закон композиции (грамматика);
$\equiv$ – операция сопоставления элементов различных систем на уровне знаков, обеспечивающая их тождество в функциональной, понятийной и терминологической системах.

[7] Данное определение отражает и относительность знаний. Наличие в определении онтологии закона композиции означает возможность существования мультионтологии – более чем одной онтологии (целостных самодостаточных «концептуализаций» ПрО), определенных на одном множестве сущностей и отношений.

[8] Примеры операций над онтологиями также приведены в [20].

[9] Рассматривается граф $G(V, E)=<V, E>$, где V – множество вершин графа, а E – множество его ребер.
Граф G обладает свойством мультиграфа, если:
$$\exists G_i(V_i,E_i) \subset G, G_j(V_j,E_j) \subset G : (\{V_i\} = \{V_j\}) \wedge (\{E_i\} \neq \{E_j\})$$
Граф G обладает свойством метаграфа, если:
$$\exists V_j \in V : (\{V_j\} \subset \{V \setminus V_j\})$$

[10] В первую очередь это относится к функциональной онтологии.

дуг, но и смысловой «направленностью», отражающей эволюцию смыслового образа объекта/результата. Это означает, что имя отдельной сущности или отношения в графе будет представлено в нескольких «экземплярах». В этом случае можно утверждать, что и понятия, и отношения выступают как лингвистические переменные, конкретный смысл которых доопределяется ситуацией – хорошо определенным контекстом[11], специфицируемым типами отношений и характером связанных сущностей. И, поскольку граф представляет *целенаправленный процесс*, должен быть определен порядок следования вершин, в том числе могут быть заданы исходные и конечные (целевые) вершины.

При этом необходимо понимать, что визуализация всего графа онтологии неконструктивна: вследствие сравнительно большой размерности и динамического характера задания контекста пользователю будет довольно трудно визуально выделять и анализировать соответствующие подграфы.

Использование онтологических представлений и операций при поиске

Семантическая сеть, представляемая графом функциональной системы онтологии предметной области, либо отдельного решения (как и статьи понятий в понятийной системе), позволяет не только учитывать роль каждого отдельного семантического элемента, но и расширить множество поисковых точек входа.

Для построения и визуализации онтологических образов документов разработаны программно-информационные средства, которые позволяют осуществлять функции:

1) перехода от вершины графа к фрагментам текста, содержащим соответствующую вершине сущность;

2) перехода от вершины графа к поиску документов по имени вершины, в том числе, с указанием контекста (с выбором окружающих вершин и связей);

3) объединения, пересечения, проекции онтологии;

4) масштабирования онтологии;

5) выделения подграфа в соответствии со специфицированным аспектом.

С точки зрения поисковых операций первая и вторая функции представляют собой реализацию тривиального отбора по ключевым словам. Однако, поскольку выбираемые термины явным образом контекстно определены через отношения со смежными понятиями (причем, с явным указанием типа связи), эти связи (точнее, их имена) могут быть явно или неявно использованы в условии отбора.

Использование функций объединения и масштабирования онтологий обеспечивает возможности сопоставления текстов, близких по смыслу, но имеющих мало общих терминов. Результирующий граф будет включать не только общие узлы функционального и понятийного уровней, но и ассоциированные семантически близкие тезаурусные понятия и связи, что в итоге позволит повысить семантическую связность сети и обеспечит большую сопоставимость. В [20, 21] приведены некоторые примеры применения этих операций для научных и технических документов.

Рассмотрим использование операции аспектной проекции, в которой аспект формально задается функциональной системой $S_a = <M_a, A_a, R_a, Z_a>$, где M_a – множество опорных понятий, A_a – множество характеристических свойств, R_a – множество функциональных отношений, характерных для данного аспекта, Z_a – правила, определяющие построение цепей (ограничивающие включение той или иной триады в состав пути на графе на основе анализа характеристических свойств соответствующих дуг и узлов).

Множество аспектов[12] в работе определяется в соответствии с моделью деятельности, рассматриваемой с системной точки зрения. На верхнем уровне модели выделяются два класса – структурные и поведенческие аспекты. Структурные включают аспекты, отражающие внутреннюю морфологию системы (взаимосвязь и состав элементов), и внешнюю – свойства, проявляющиеся при взаимодействии, в частности, с окружением системы (в том числе и качественные оценки). Примеры структурных аспектов – «Состав», «Следование», «Соответствие», «Форма», «Изменение», «Внешние признаки» и т.п. Поведенческие аспекты отражают зависимости и соотношения, определяемые, в том числе, обобщенной функциональной моделью деятельности. Примеры поведенческих аспектов – «Эффективность», «Управляемость», «Обеспеченность» и т.п.

Аспекты также могут отражать и характер представления смысла, позволяя анализировать разнообразие и интенсивность использования отдельного понятия (формируя соответствующий гиперграф) или его эволюцию (формируя направленный подграф, в котором понятие будет представлено в «исторической» последовательности его употребления[13] в тексте и, соответственно, в разных контекстах – связанным отношениями разного типа с разными понятиями).

При информационном поиске, в общем случае, используется *таксономия аспектов* – часть его лингвистического обеспечения. Такая таксономия (являясь объектом, открытым для расширения и модификации) задает множество возможных аспектов, каждый из которых фиксирует связь аспекта с классами характеристических отношений. Классы отно-

[11] Это также управляемый контекст: через задание аспекта и/или параметр концептуальной глубины и/или широты.

[12] Аспектные представления являются одной из методологических основ синтеза нового знания [22]. В основе этой модели синтеза знаний как самоорганизующегося процесса лежит структурная особенность системы – сложная система может быть описана при помощи набора относительно независимых аспектных представлений. В процессе декомпозиции не только выделяются и связываются составляющие, но и формируется *схема декомпозиции* – система характеристических признаков, в соответствии с которой и проводится декомпозиция.

[13] Здесь становится очевидной неоднозначность отображения (и, соответственно, идентификации вершин): вершины, по существу, должны иметь одинаковое имя (самого понятия), однако содержание будет разным, поскольку смысл понятия определяется и контекстом, образуемым связанными понятиями и типом отношений.

шений, в свою очередь, связаны с лингвистическими конструкциями[14], которые представляют собой конкретные отношения, раскрывающие аспект в тексте.

Аспект задается функциональной системой $S_a = <M_a, Ø, R_a, Ø>$, где M_a – множество узлов графа онтологии, R_a – множество классов отношений для данного аспекта (приводится к множеству дуг графа через аспектную таксономию), множество характеристических свойств и множество правил формирования маршрутов пусты ($A_a = Ø$, $Z_a = Ø$).

Фрагмент графа онтологии (точнее, ее функциональной системы), автоматически построенной по тексту раздела документа «*Общее описание энергоблока*», представлен на рис. 1.

Примеры результатов применения операции аспектной проекции представлены на рис. 2*а* и 2*б*. На рис. 2*а* изображена проекция сущности «вода» (M_a = {«вода»}) в аспекте «Изменение». В таксономии, которая была использована для построения примера, аспект «Изменение» раскрывается классами отношений «Изменение состояния» (лингвистические конструкции – «… охлаждать …», «…гореть…», «…ускорять…» и т.д.), «Изменение состава» (лингвистические конструкции – «… обессолить …», «… химически очистить …», «…заполнить…», «…добавить…» и т.д.), «Быть целью / предназначением» (лингвистические конструкции – «… предназначить для …», «…выполнить для…», «…служить…» и т.д.), «Быть средством / инструментом» (лингвистические конструкции – «… поступать по …», «…использовать…», «…эксплуатировать…» и т.д.), «Перемещение» (лингвистические конструкции – «… отводиться по …», «…бегать…», «…бросать…», «…распространять…» и т.д.) и т.д.

Аспектная проекция показала три основных состояния воды – охлажденное, обессоленное и химически очищенное, – с каждым из которых связано свое множество сущностей.

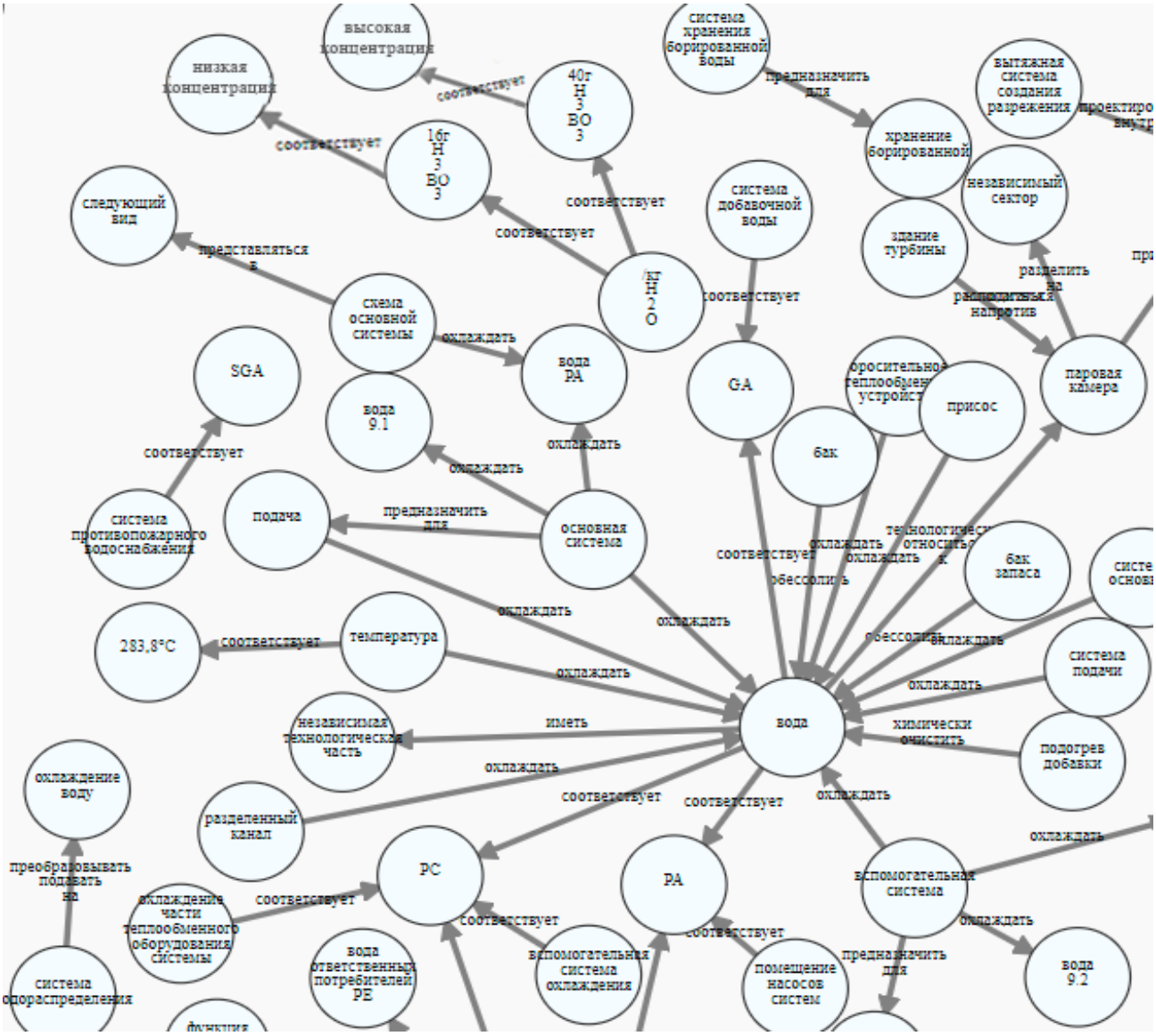


Рис. 1. Фрагмент графа онтологии, построенный по тексту раздела документа «Общее описание энергоблока»

[14] Соответствие отношений и лингвистических конструкций приведено в [3].

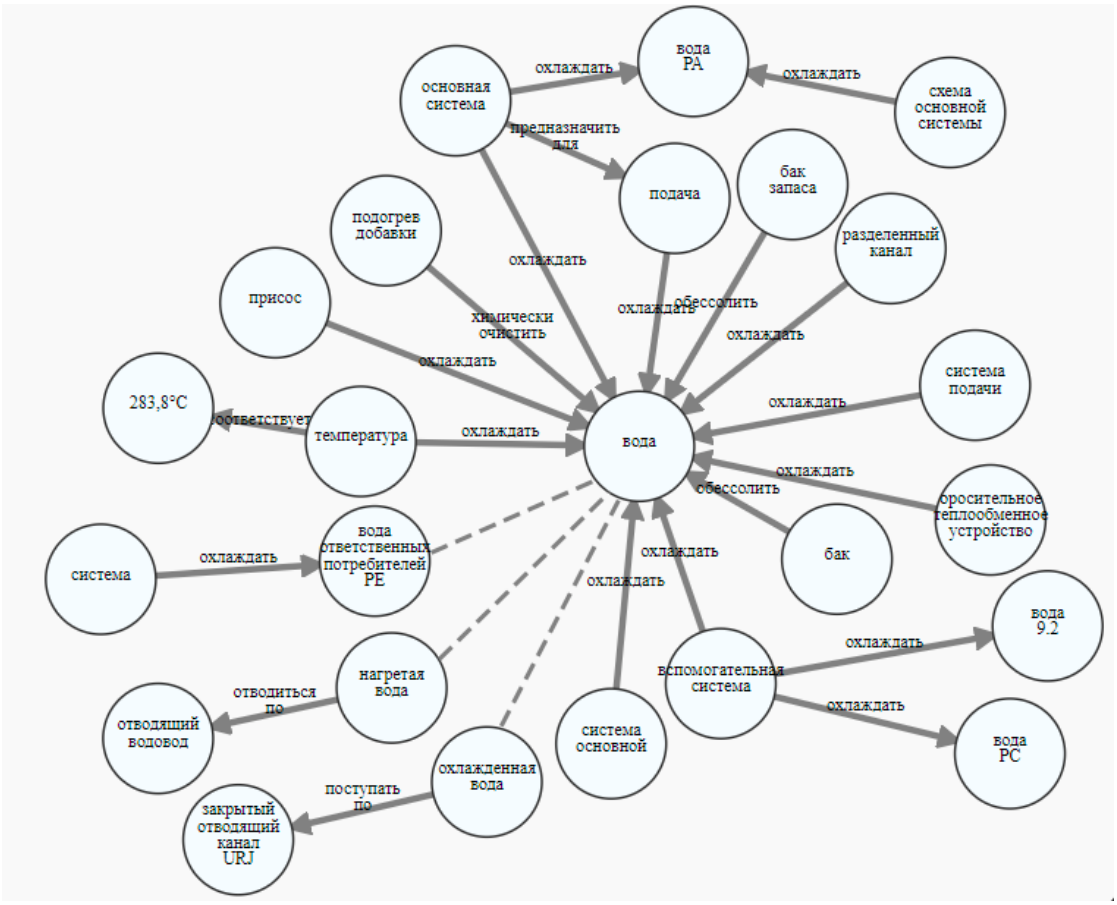


Рис. 2*а*. Проекция графа в аспекте «Изменение» для сущности «Вода»

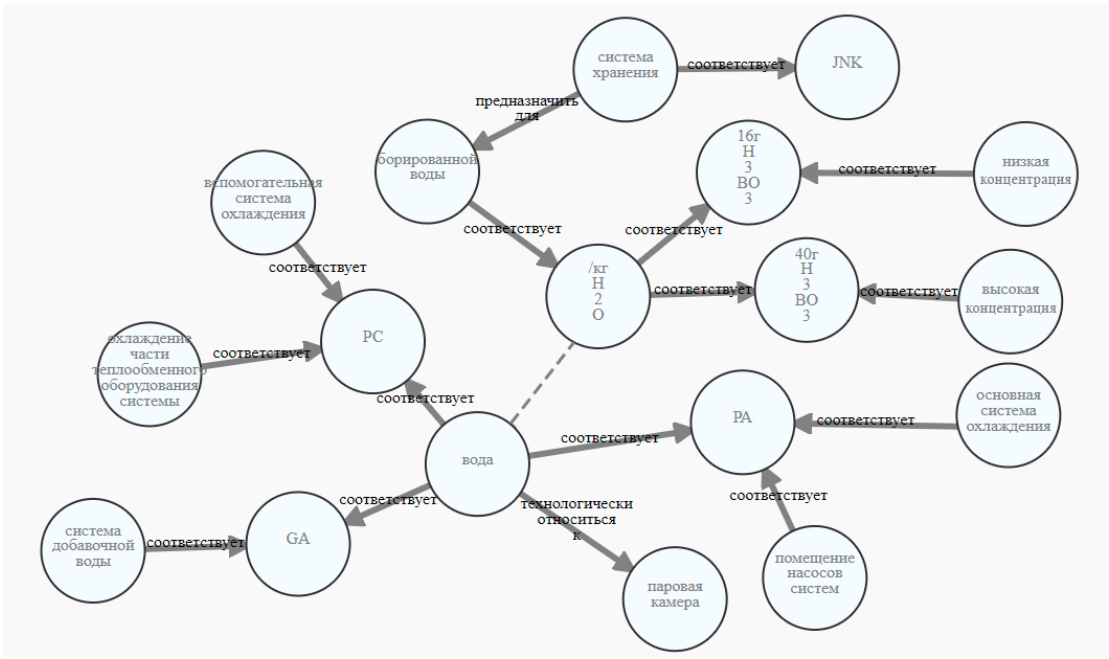


Рис. 2*б*. Проекция графа в аспекте «Соответствие» для сущности «Вода».

Рис. 2*б* иллюстрирует проекцию исходной онтологии в аспекте «Соответствие». Аспект задан в таксономии следующими классами характеристических отношений: «Быть основанием» (лингвистические конструкции – «… технологически относиться к …», «…базироваться на…», «…доказывать…» и т.д.), «Быть условием» (лингвистические конструкции – «…выделить при…», «…предусматривать…», «…рассчитывать из…» и т.д.), «Тождество» (лингвистические конструкции – «… соответствует …»,

«…оценить…», «…удовлетворять …» и т.д.), «Быть целью / предназначением» (лингвистические конструкции – «… предназначить для …», «…выполнить для…», «…служить…» и т.д.), «Быть средством / инструментом» (лингвистические конструкции – «…назначать…», «…посредством…», «…применять…» и т.д.) и т.д. Такая проекция, как видно на рис. 2*б*, позволяет связать аббревиатуры с их расшифровкой (*PC* – вспомогательная система охлаждения, *GA* – системы добавочной воды).

Приведенные примеры иллюстрируют предлагаемую в настоящей работе методику редукции текста, которая основана на глубоком семантическом индексировании с использованием таксономии классов отношений. Семантический образ документа в рамках этой методики представляет собой граф, в узлах которого размещены имена сущностей, а дуги отражают функциональные отношения. Необходимо отметить, что качество построения такого образа в значительной мере определяется характеристиками применяемых методов семантического анализа текстов: 1) результаты графематического анализа во многом зависят от наличия разного рода ошибок в исходном тексте (например, вершина «/кг H_2O» на рис. 2*б* выделена как таковая в следствие разрыва строки предложения); 2) морфологический анализ требует полноты словарей морфологических признаков, в том числе и для специфических терминов предметной области; 3) семантико-синтаксический анализ может использовать разные способы формирования имен сущностей и отношений (например, существенным параметром при формировании словосочетаний является максимальное допустимое количество слов и наличие семантических словарей, а выделение триплетов в тексте зависит от применяемых лексико-семантических шаблонов и т.п.).

И отметим, что в данной конкретной программной реализации операции проекции не использовались лингвистические компоненты знаковой и понятийной систем онтологии (т.е. операция применялась только к графу функциональной системы), вследствие чего автоматически не были выделены связи эквивалентности (например, синонимия терминов «вода» и «H_2O») и родо-видовые связи (например, «вода» является более широким понятием по отношению к понятию «охлажденная вода»), заданные в понятийной и знаковой системах (на рис. 2*а* и 2*б* эти связи указаны пунктиром).

Использование процессных онтологий как средства навигации при поиске[15]

Эффективное управление предприятиями и процессами производственной деятельности – сложная проблема, требующая решения взаимосвязанных и взаимосогласованных задач организации работ, планирования, мониторинга и контроля исполнения, а также развития. И онтологии здесь представляются удобным поисковым инструментом, поскольку в целом отражают знания о производственных процессах и созидательной деятельности людей. Так, вопрос использования онтологий для решения задач ситуационного управления предприятиями в режиме реального времени рассматривается в [23].

Любая деятельность в рамках любого проекта может рассматриваться как последовательность этапов, на каждом из которых решаются свои задачи. В процессе этого решения могут использоваться как документы, созданные на предыдущих этапах, так и нормативные документы (государственные, отраслевые стандарты, рекомендации и т.п.), регламентирующие состав и форму представления работ. Результат решения также фиксируется в виде документа (отчета или иной соответствующей этапу форме). Порядок и состав работ этапа проекта может определяться в рамках одного нормативного документа, но детализироваться в рамках других документов. Кроме того, состав работ может зависеть от конкретных условий или результатов, полученных на предыдущих этапах. Необходимость соблюдения условий для перехода между этапами и разрозненность (а иногда и противоречивость) описания самих условий может привести к формальной блокировке работ по проекту, поскольку нужные условия в принципе не смогут быть достигнуты. Кроме того, надо отметить, что фактор формы документации (существенно влияющий на характер и полноту изложения и, соответственно, понимания) практически не учитывается при выполнении информационного поиска.

В этих условиях представляется конструктивным в качестве дополнительного средства поиска использовать онтологии организационных процессов, которые будут играть роль предопределенной потенциальной траектории расширения поискового пространства. Граф онтологии может использоваться как некоторый шаблон тем поискового запроса, *смежных в рамках жизненного цикла* проекта. При этом выбор подмножества нормативных документов позволяет достичь нужной полноты описания, поскольку разные виды документов представляют существо объекта с разной полнотой и детальностью.

Такие онтологии, представляющие априорную информацию, предварительно строятся на основе текстов нормативных документов, определяющих порядок выполнения работ (научно-исследовательские работы – НИР, опытно-конструкторские работы – ОКР и т.д.). Далее в процессе поиска представление запроса в виде графа (в общем случае обладающего свойствами мульти- и метаграфа), где отдельная вершина или дуга может рассматриваться как точка входа в информационное пространство (множество всех доступных документов по проекту и/или вне его), позволяет раскрыть содержание этапа как в плане организации работ (на основе онтологии ассоциированных нормативных документов), так и в плане их существа (при объединении с онтологиями соответствующих этапу документов, представляющих результаты работ или используемых при их выполнении).

Состав работ и объектов при выполнении конкретного проекта (по отношению к стандартам) всегда уточняется, дополняется, что фиксируется в техническом задании (ТЗ) и другой документации проекта. В этом случае операции над онтологиями обеспечивают возможность и для выявления формальными методами соответствий/расхождений между сравниваемыми документами.

[15] Практический материал подготовлен студентом магистратуры НИЯУ «МИФИ» Т.Г. Исмаиловым

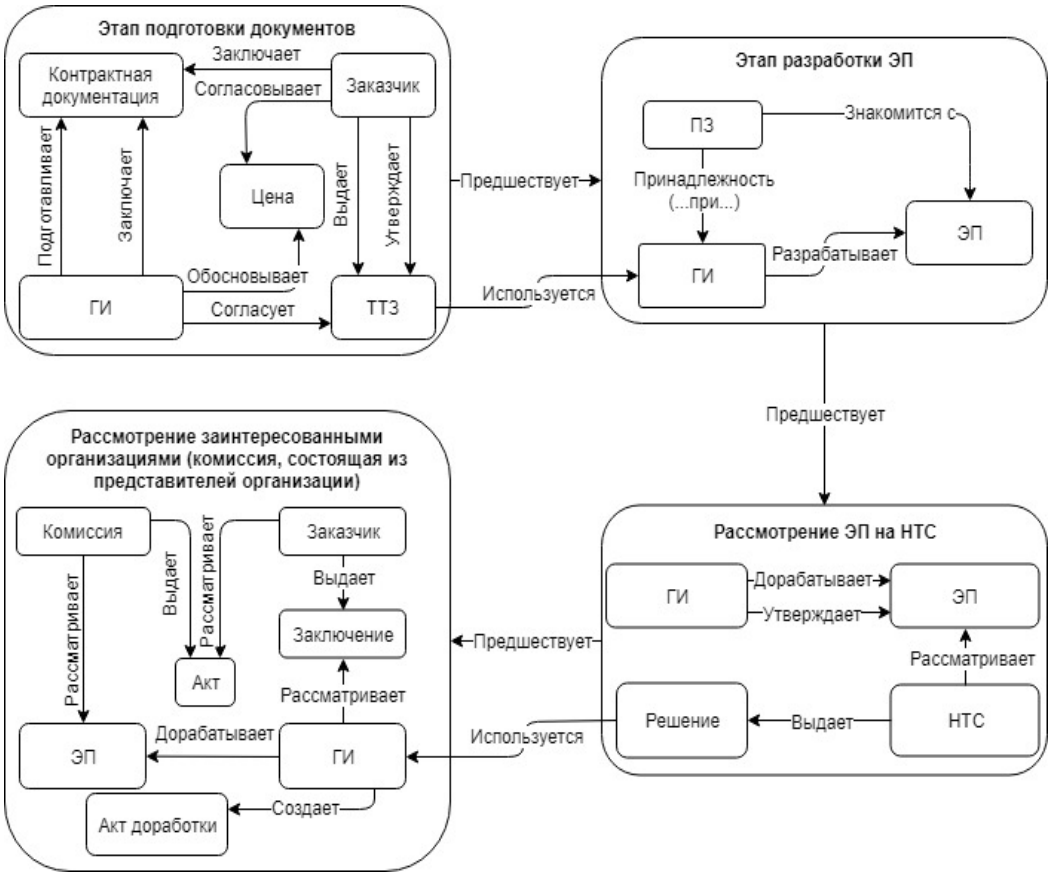


Рис. 3. Фрагмент метаграфа, построенного по тексту ГОСТ РВ 15.110-2003, и описывающего структурную и процессную составляющие работ, выполняемых в рамках ОКР
(ГИ – Головной исполнитель; ЭП – Эскизный проект; НТС – Научно-технический совет;
ПЗ – Представитель заказчика)

Проиллюстрируем это на примере. На рис. 3 представлен фрагмент графа онтологии, построенной[16] на основе текста ГОСТ РВ 15.203-2001 «Военная техника. Порядок выполнения опытно-конструкторских работ по созданию изделий и их составных частей. Основные положения». Атомарные вершины графа содержат имена объектов и субъектов ОКР (наименования документов, имена участников работ и т.п.), а этапы работ представлены метавершинами. Подграф метавершины включает (с точки зрения формулировки поискового запроса) лексику определенного этапа и может быть использован для поиска как нормативных документов, так и документов реальных НИР и ОКР и т.д.

Поиск нормативных документов в качестве результата может выдать, например, граф конкретного стандарта из базы данных онтологий, подготовленных экспертным путем, или непосредственно текст стандарта и его автоматически сформированный онтологический образ. На рис. 4 (слева) представлен граф онтологии, подготовленный по ГОСТ 15.016-2016 «Система разработки и постановки продукции на производство (СРПП). Техническое задание. Требования к содержанию и оформлению». Такой граф может служить основой, в частности, для сопоставления конкретных документов с требованиями стандартов.

Например, вершина «ТТЗ» (тактико-техническое задание) в метавершине «Этап подготовки документов» (см. рис. 3) может раскрываться подграфом, построенным уже по тексту ГОСТ 15.016-2016 (рис. 4, слева) или же по тексту ТЗ реального проекта (на рис. 4 справа представлен фрагмент описания реального ТЗ на разработку автоматизированной информационной системы – АИС). Раскрытие вершины может происходить следующим образом: пользователь, выделяя вершину, передает связанный с ней термин (и, возможно, некоторые другие уточняющие термины) как запрос для поиска в ассоциированных базах данных. Полученные результаты возвращаются пользователю в виде графа или ссылок на релевантные документы / фрагменты текста.

[16] Онтология строится автоматически, однако, на рис. 3 представлено обработанное вручную изображение графа, где вручную выделены метавершины (средства отображения и операций над метаграфами в составе АИС *xIRBIS* находятся в стадии разработки).

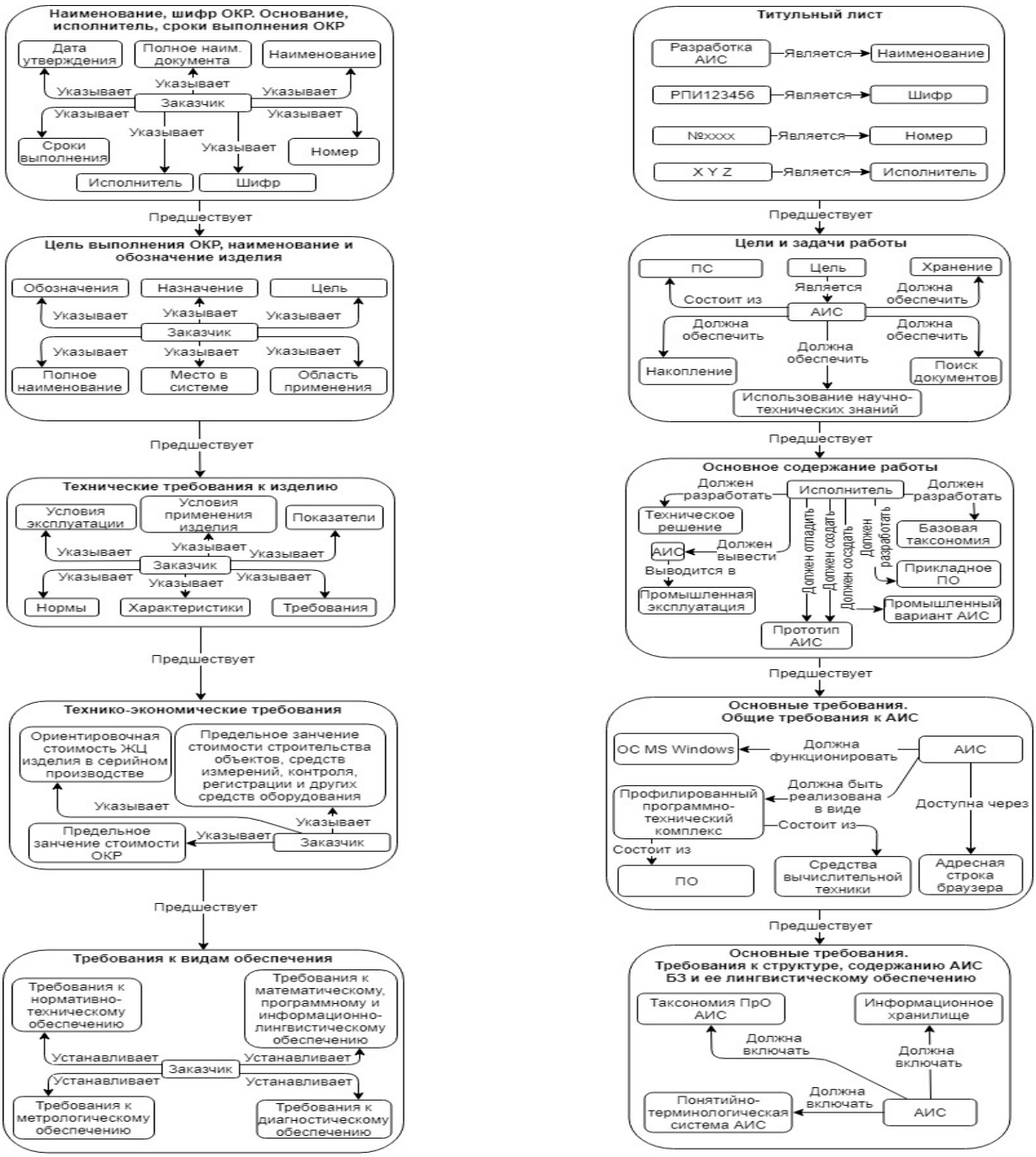


Рис. 4. Соотнесение подграфов «Техническое задание»: построенного на основе текста ГОСТ 15.016-2016 – слева, и построенного на основе текста реального технического задания – справа (здесь *ПО* – Программное обеспечение; *ПС* – Программные средства)

В каком-то смысле онтологии являются «контурной картой», на которой пользователь реализует траекторию поиска: рациональный порядок использования ключевых понятий для поиска релевантной информации обеспечивается за счет визуально доступных возможных путей между рассматриваемыми вершинами. На примере перехода от вершины «ТТЗ» (рис. 3) к графу на рис. 4 (слева), представляющему требования к ТТЗ (построенному по тексту ГОСТ 15.016-2016), это можно представить следующими поисковыми действиями: по запросу «ТТЗ» получаем документы и, соответственно, графы онтологий. Для того чтобы выделить только те фрагменты, которые интересуют пользователя, можно воспользоваться операцией аспектной проекции, т. е., выбирая вершину «ТТЗ» как исходную точку, отбираем только вершины, связанные с ней, а полученные вершины сами становятся новыми запросами.

Онтологический подход к представлению семантики документов позволяет проводить и сравнение документов, используя операции объединения, пересечения графов и т.д. Так, визуальное сравнение графов (образов документов), изображенных на рис. 4, показывает, что метавершина «Цель выполнения ОКР…» (рис. 4, слева) является структурной основой для метавершины «Цели и задачи работы» (рис. 4, справа). Конструкция, состоящая из вершин «Цель» и «Заказчик», связанных ребром «Указывает», соотно-

сится с вершинами «Цель» и «АИС», связанных ребром «Является» из метавершины «Цели и задачи работы», так как имеются понятийно тождественные вершины («Цель») и принадлежащие одному классу отношений связи. Вершина «Назначение» косвенно отождествляется с группой вершин и отношений из метавершины «Цели и задачи работы»: «АИС – должна обеспечить – Хранение», «АИС – должна обеспечить – Поиск документов», «АИС – должна обеспечить – Накопление». Вершина «Требования» из метавершины «Технические требования к изделию» (рис. 4, слева) в нормативном документе раскрывается через метавершину «Основные требования. Технические требования к АИС» (рис. 4, справа) графа онтологии реального ТЗ. На рис. 4 справа (по тексту реального ТЗ) отсутствуют регламентированные в ГОСТ 15.016-2016 такие составляющие как, «Характеристики», «Нормы», «Место в системе» и др., а также отсутствует часть «Технико-экономические требования».

Таким образом, использование процессных онтологий позволяет узнать не только о существующих работах, этапах, порядке их выполнения, субъектах деятельности, документах, соответствующих конкретным работам, но и место определенного документа в жизненном цикле проекта, а методы операций над графами позволяют проводить оценку соответствия проектной документации формальным требованиям.

ЗАКЛЮЧЕНИЕ

Управляемая визуализация онтологий дает возможность в компактной форме представить понятия предметной области и связи между ними, анализировать смысловую окрестность понятия. Это позволяет выделять и наиболее значимые для когнитивной деятельности понятия, которые могут быть использованы для продолжения поиска, в частности путем перехода по гипертекстовой ссылке (как свойству вершины). Применяя к графам операции над онтологиями, можно строить и визуализировать новые (неявные) пути между вершинами, т.е. выявлять неявные связи между объектами предметной области.

Возможность развития поискового процесса (формулирования и реформулирования запроса) основывается на глубоком многоаспектном полнотекстовом индексировании документа (что позволяет увеличить количество точек входа, в том числе и альтернативных интерпретаций возможных комбинаций понятий) и на использовании операции аспектных проекций, позволяющих сократить количество операционных объектов (сведение анализа большого целого к последовательности однофакторных задач анализа объектов приемлемой размерности).

Онтология, как высокоинформативная форма представления семантики документа, дающая представление о роли и значимости терминов и связей, предъявляя их в контексте конкретной ситуации. Онтология, построенная, например, на множестве релевантных теме поиска документов, позволяет локализовать взгляд на проблему, составляющую объект или предмет поиска. Более того, по характеру динамики использования понятий, связей и их контекстов во времени можно будет судить и об изменении видения предметной области. Онтология позволяет также проводить поиск как внутри документа, выделяя цепочки понятий, связанных в заданном аспекте, так и обеспечивает информационную «навигацию» по предметной области путем построения объединенной онтологии и/или проекции. В целом всё это создает основу для понимания пользователем, почему был выдан тот или иной документ, а, главное, в каком направлении развивать поиск – какие понятия в каком семантическом окружение надо использовать для развития запроса. А применение графовых представлений позволяет перейти к графически управляемым навигационным технологиям экспертного поиска: визуализация объектов, с одной стороны, фокусирует внимание, а с другой – увеличивает и упорядочивает разнообразие точек входа, в том числе, за счет улучшения «видимости» ключевых аспектов документа. В итоге такой подход создает основу для исследования нескольких гипотез .

СПИСОК ЛИТЕРАТУРЫ

1. Голицына О.Л., Максимов Н.В., Окропишина О.В., Строгонов В.И. Онтологический подход к идентификации информации в задачах документального поиска // Научно-техническая информация. Сер. 2. – 2012. – № 5. – С. 1-10; Golitsyna O.L., Maksimov N.V., Okropishina O.V., Strogonov V.I. The ontological approach to the identification of information in tasks of document retrieval // Automatic Documentation and Mathematical Linguistics. – 2012. – Vol. 46, №3. – P. 125-132.
2. Максимов Н.В. Методологические основы онтологического моделирования документальной информации // Научно-техническая информация. Сер. 2. – 2018. – № 3. – С. 6-22; Maksimov N.V. The Methodological Basis of Ontological Documentary Information Modeling // Automatic Documentation and Mathematical Linguistics. – 2018. – Vol. 52, №2. – P. 57–72.
3. Максимов Н.В., Гаврилкина А.С., Андронова В.В., Тазиева И.А. Систематизация и идентификация семантических отношений в онтологиях научно-технических предметных областей // Научно-техническая информация. Сер. 2. – 2018. – №11. – С. 32-42; Maksimov N.V., Gavrilkina A.S., Andronova V.V., Tazieva I.A. Systematization and Identification of Semantic Relations in Ontologies for Scientific and Technical Subject Areas // Automatic Documentation and Mathematical Linguistics. – 2018. – Vol. 52, №6. – P. 306–317.
4. Hoeber O., Yang X., Yao Y. Conceptual query expansion // Proc. of the Third Int. Conf. on Advances in Web Intelligence (June 6-9, 2005, Lodz, Poland). – Berlin, Heidelberg: Springer. – P. 190–196.
5. Hoeber O., Yang X., Yao Y. Visualization Support for Interactive Query Refinement // Proc. of the 2005 IEEE/WIC/ACM Int. Conf. on Web Intelligence (September 19-22, 2005, Compiegne, France). – Washington, DC, USA: IEEE Computer Society. – P. 19–22.

6. Gorbun E.S., Maksimov N.V., Monankov K.V., Nizametdinov Sh.U. Model and tools for interactive analysis of dynamics and relations of scientific information publications flows // Scientific Visualization. – 2015. – Vol. 7(5). – P. 12–25.
7. Hoeber O. Information Visualization for Interactive Information Retrieval // Proc. of the 2018 Conference on Human Information Interaction & Retrieval (March 11–15, 2018, New Brunswick, NJ, USA). – New York, NY, USA: ACM. – P. 371–374.
8. Steve K. Don't Make Me Think, A Common Sense Approach to Web Usability. – Indianapolis, USA: New Riders, 2000. – 216 p.
9. Cooper A., Reimann R., Cronin D., Noessel C. About Face: The Essentials of Interaction Design 4th Edition. – Indianapolis, USA: John Wiley & Sons, Inc., 2014. – 720 p.
10. Rose D., Levinson D. Understanding user goals in web search // Proc. of the 13th Int. Conf. on World Wide Web (May 17-22, 2004, New York City, USA). – New York, NY, USA: ACM. – P. 13-19
11. Teevan J., Dumais S., Horvitz E. Personalizing Search via Automated Analysis of Interests and Activities // Proc. of SIGIR (August 15-19, 2005, Salvador, Brazil). – New York, NY, USA: ACM. – P. 449–456.
12. Hoeber O., Yang X. Interactive Web Information Retrieval Using WordBars // Proc. of the 2006 IEEE/WIC/ACM Int. Conf. on Web Intelligence (December 18-22, 2006, Hong Kong, China). – Washington, DC, USA: IEEE Computer Society. – P. 875-882.
13. IBM i2 Analyst's Notebook. – URL: https://www.ibm.com/ru-ru/marketplace/analysts-notebook (дата обращения 04.04.2019).
14. Products built for a purpose. – URL: https://www.palantir.com (дата обращения 04.04.2019).
15. RCO Fact Extractor SDK. – URL: http://www.rco.ru/?page_id=3554 (дата обращения 04.04.2019).
16. Compreno. – URL: https://www.abbyy.com/ru-ru/science/technologies/compreno/ (дата обращения 04.04.2019).
17. Miller G.A. The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information // The Psychological Review. – 1956. – Vol. 63. – P. 81–97.
18. Максимов Н.В. Информация и знания: природа, концептуальная модель // Научно-техническая информация. Сер. 2. – 2010. – №7. – С. 1-10; Maksimov N.V. Information and knowledge: Nature and the conceptual model // Automatic Documentation and Mathematical Linguistics. – 2010. – Vol. 44, №4. – P. 177–186.
19. Golitsyna O.L., Maksimov N.V., Okropishina O.V., Okropishin A.E. Semantic Identification of Text in the Class of Tasks of Information Retrieval // Proc. of ICAI'14, WORLDCOMP'14 (July 21-24, 2014, Las Vegas, Nevada, USA). – CRSEA Press, USA. – Vol. I. – P. 47-52.
20. Голицына О.Л., Максимов Н.В., Окропишина О.В., Строгонов В.И. Онтологический подход к идентификации информации в задачах документального поиска: практическое применение // Научно-техническая информация. Сер. 2. – 2013. – №3. – С. 1-8; Golitsyna O.L., Maksimov N.V., Okropishina O.V., Strogonov V.I. An ontological approach to information identification in tasks of document retrieval: A practical application // Automatic Documentation and Mathematical Linguistics. – 2013. – Vol. 47, №2. – P. 45-51.
21. Максимов Н.В., Голицына О.Л., Ганченкова М.Г., Санатов Д.В., Разумов А.В. Семантическое ядро цифровой платформы // Онтология проектирования. – 2018. – Т. 8, №3(29). – С.412-426.
22. Урманцев Ю.А. Общая теория систем: Состояние, приложения и перспективы развития // Сб. «Система, Симметрия, Гармония». – М.: Мысль, 1988. – С.38-124.
23. Скобелев П.О. Онтологии деятельности для ситуационного управления предприятиями в реальном времени // Онтология проектирования. – 2012. – №1(3). – С.6-38.

Материал поступил в редакцию 07.05.19

Сведения об авторах

МАКСИМОВ Николай Вениаминович – профессор, доктор технических наук, профессор института Интеллектуальных кибернетических систем Национального исследовательского ядерного университета «МИФИ».
e-mail: nv-maks@yandex.ru

ГОЛИЦЫНА Ольга Леонидовна – доцент, кандидат технических наук, доцент института Интеллектуальных кибернетических систем Национального исследовательского ядерного университета «МИФИ».
E-mail: olgolitsina@yandex.ru

МОНАНКОВ Кирилл Викторович – аспирант кафедры 75 Национального исследовательского ядерного университета «МИФИ».
e-mail: kmonankov@yandex.ru

ЛЕБЕДЕВ Александр Анатольевич – ведущий математик кафедры 75 Национального исследовательского ядерного университета «МИФИ».
e-mail: lebedevalex@live.ru

БАЛЬ Наталья Александровна – студент 4-го курса Национального исследовательского ядерного университета «МИФИ» факультета бизнес-информатики и управления комплексными системами.
e-mail: baln757@mail.ru

КЮРЧЕВА София Геннадьевна – студент 4-го курса Национального исследовательского ядерного университета «МИФИ» факультета бизнес-информатики и управления комплексными системами.
e-mail: kyurcheva@yandex.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’366

В.П. Фесенко

Винительный падеж как маркер определенности объекта при переходном глаголе с отрицанием (на примере терминов родства)

Рассматривается проблема выбора родительного или винительного падежа объекта при переходном глаголе с отрицанием в диахроническом аспекте: с XVIII в. по начало XXI в. Цель – определить, в какой период и в какой мере на выбор падежа влиял фактор референтности объекта при переходном глаголе. В качестве материала для исследования выбраны примеры из основного корпуса Национального корпуса русского языка с презумптивным объектом – названиями родства по женской линии.

Ключевые слова: *отрицание, родительный падеж, винительный падеж, определенность, презумптивные существительные*

ВВЕДЕНИЕ

В современной русской грамматике вопрос о выборе родительного / винительного падежа объекта при отрицании относится к проблемам, изучению которых посвящены десятки исследований. За два столетия было выдвинуто множество гипотез о том, какие факторы влияют на выбор падежа при отрицании. Среди них предпочтение отдается влиянию скрытых семантических категорий, в частности – категории определенности/неопределенности ([1–4] и др.). В «Русской грамматике» (1980) читаем: «Употребление вин.п. со знач. определенного объекта создает противопоставление: "вин.п. для выражения определенного объекта – род.п. для выражения неопределенного объекта": *Он не получил письмо* (определенное, известное) – *Он не получил письма* (неопределенного, какого-либо)» [5, с. 418]. Е.В. Падучева, признавая значимость этого фактора, интерпретирует его в терминах теории референции и соединяет на его основе два правила – выбор падежа при отрицании в бытийных предложениях и выбор падежа при отрицании в словосочетании с переходным глаголом: «Подтвердилась референциальная основа семантики генитива. Эта основа для объекта и для субъекта одна и та же: генитив объекта, как и генитив субъекта, в случае индивидного терма выражает либо нереферентность, либо наблюдаемое отсутствие» [4, с. 41]. Таким образом, определенность/неопределенность можно считать ключевым фактором при выборе падежа объекта при отрицании.

Однако в первой половине XIX в. фактор определенности/неопределенности не действовал, вместо него выбором падежа руководило правило грамматики: «…Если оное [действие. – *В.Ф.*] выражается съ отрицаніемъ, то сей падежъ замѣняется *родительнымъ*; напримѣръ: *...не люблю ссоры*» [6, с. 269]. Другими словами, в этот период была важна только синтаксическая конструкция.

Позже требование грамматики стало уступать семантическим нюансам, передаваемым с помощью винительного или родительного падежа. В частности, к таким семантическим аспектам относится и значение определенности/неопределенности объекта. Вопрос заключается в следующем: с какого момента семантика становится более важной в письменной речи, чем грамматическое правило? Ответ может дать рассмотрение действия определенности объекта при отрицании в диахронии.

КРИТЕРИИ СОСТАВЛЕНИЯ ВЫБОРКИ

Для исследования категории определенности/неопределенности выбраны примеры с объектами – названиями родства по женской линии в единственном числе. Выбор обусловлен следующим:

- определенность названий родства связана с тем, что это имена презумптивные (обычно говорят об имеющемся родственнике), поэтому винительный падеж очень вероятен даже при действующем в грамматике правиле: «Значение “определенности” у имен типа *муж, жена, теща, свекровь* и т. п. связа-

но… с “презумпцией условной единственности”, состоящей в том, что если существует хотя бы один объект, подводимый под данную категорию, то не существует никакого другого объекта, который можно было бы подвести под ту же категорию» [7, с. 250];

- названия родства можно включить в состав эгоцентрической лексики (словам, ориентированным на говорящего), поскольку они относятся к личной сфере субъекта высказывания;
- презумптивность и эгоцентричность названий родства обусловливают тот факт, что эта группа лексики занимает крайнюю позицию в противопоставлении «определенное – неопределенное», следовательно, работа только с этой группой даст чистый результат;
- имена женского рода в единственном числе не создают омонимию родительного и винительного падежей, которая есть у названий родственников мужского пола (*не вижу отца* – родительный или винительный?) и у названий родственниц во множественном числе (*не вижу сестер* – родительный или винительный?).

Для работы с названиями родства необходимо провести параллель между различными группами терминов, используемыми в разных исследованиях при обозначении интересующего нас фактора.

Контексты с названиями родства могут быть трех типов.

1. Относящиеся к конкретному денотату, находящемуся, условно говоря, в поле зрения субъекта и занимающему определенную точку во времени и пространстве по отношению к субъекту:

(1) *Он приехал внезапно, без предупреждений, чтобы, видимо,* ***не волновать мать****, но вышло еще хуже: от счастья она потеряла дар речи.* [Руслан Киреев. Шишига // Библиотека «Огонек», 1989].

2. Относящиеся к единичному денотату, существующему или существовавшему в мире, но не входящему в поле зрения или сознания субъекта, без привязки к конкретной точке во времени и пространстве:

(2) *А сиротам никто* ***не заменит родной матери****…* [И.А. Бунин. Лирник Родион (1913)].

3. Не относящиеся ни к какому конкретному денотату и употребленные в обобщенном времени[1]:

(3) *Скажу больше: человек, который* ***не любит мать****, давшую ему жизнь, выкормившую и воспитавшую его,* ***—*** *такой человек мне лично подозрителен.* [Л.И. Брежнев. Жизнь по заводскому гудку (1980)].

Этим типам контекстов соответствуют три группы терминов (см. таблицу): по «Русской грамматике» под ред. Н.Ю. Шведовой (РГ-80) [5]; по теории референции, используемой в работах Е.В. Падучевой; термины «суперкатегории», предложенные Т.В. Булыгиной и А.Д. Шмелевым [8, 9].

Наиболее общие термины, используемые в РГ-80, – «определенность» и «неопределенность» – для нашего исследования не подходят, так как не отражают разницы между первым и вторым типами употреблений. Внутренняя форма терминов теории референции указывает на связь между первой и второй группой как референтными и противопоставляет их нереферентным. В терминологии Т.В. Булыгиной и А.Д. Шмелева этого противопоставления нет, поэтому для настоящего исследования подходят термины именно «суперкатегории». Кроме того, толкования терминов «суперкатегории» учитывают привязку ко времени и пространству:

- «инстанты», или «инстанции» – «индивид в конкретной пространственно-временной локализации» [8, с. 51];
- абстрактные индивиды – закрытые множества объектов, взятые в отвлечении от пространственно-временной локализации;
- классы – «открытые, неперечислимые множества объектов» [9, с. 56], действующие в обобщенном времени.

Привязка ко времени–пространству, которая есть в этих трех группах терминов, важна для интерпретации примеров с названиями родства. Так, в примере (1) – это определенная мать, существующая и действующая в конкретной точке времени и пространства, взятой относительно к расположению субъекта предложения. В примере (2) у каждого из субъектов (сирот) есть определенная мать, но эти объекты не привязаны к конкретной точке времени и пространства – так же, как не привязаны к ним субъекты. Пример (3) – в обобщенном времени; каждый человек, соответствующий приведенному описанию, может стать субъектом предложения.

1.	РГ-80	определенность		неопределенность
2.	Теория референции Е.В. Падучевой	конкретная референтность	определенная референтность	нереферентность
3.	«Суперкатегория» Т.В. Булыгиной и А.Д. Шмелева	«инстанции»	абстрактные индивиды	классы

[1] В генеритивном регистре по коммуникативной грамматике Г.А. Золотовой.

Зачастую в полученной выборке трудно отнести тот или иной пример к классу или к абстрактному индивиду. Так, в примере (2) подразумевается, что у каждого сироты есть родная мать, соответственно именная группа *родная мать* представляет собой закрытое множество – по количеству объектов в связанной с ней именной группе *сироты*, – взятое в отвлечении от места и времени. Однако это множество введено в генерализованный (обобщенно-временной) контекст – как в случае с именами классов. Согласно А.Д. Шмелеву, у классов и абстрактных индивидов действительно много общего: «Из трех выделенных разновидностей референтов – классов, абстрактных индивидов и "инстантов" – первые две… обнаруживают значительное сходство с точки зрения сочетаемости с ПВ [предикатными выражениями – *В.Ф.*] и вместе противопоставлены "инстантам"» [9, с. 61]. Чтобы избежать путаницы, в данном исследовании абстрактные индивиды и классы рассматриваются в одной группе под общим названием «имена класса». Таким образом, мы исходим из того, что в группе имен «инстантов» все примеры – с референтным объектом, относящимся к конкретному денотату, расположенному в определенной точке времени-пространства. Доминирование винительного падежа в группе «инстантов» и родительного – в именах класса будет свидетельствовать о действии фактора определенности/неопределенности.

Анализируемые нами примеры (всего 1404 примера) взяты из основного корпуса Национального корпуса русского языка (НКРЯ)[2], охватывающего период с 1800 по 2017 гг. Формула поиска:

"не
+ (1;1) V, indic, tran
+ (-2;2) S, (gen|gen2|acc|acc2), sg".

Далее мы представим результаты корпусного исследования по столетиям. В каждом столетии отдельно рассмотрим «инстанты» как чистую репрезентацию референциальной силы названий родства и проведем их сравнение с именами класса.

ИЗМЕНЕНИЕ ДОЛИ ВИНИТЕЛЬНОГО ПАДЕЖА в XVIII-XXI вв.

XVIII век. Всего в этом столетии[3] найдено 24 примера с названиями родства в отрицательной конструкции (17 – с «инстантами» и 7 – с именами класса). Из них: 6 примеров с объектом *мать*, 9 – *дочь*, 4 – *жена*, 5 – *невеста*. Примеров с другими названиями родства за этот период в основном корпусе не найдено. Все примеры единичны, поэтому о закономерности в них можно говорить лишь условно.

В этот период фактор референтности очевиден только в примерах со словом *жена*. Примеры с этим объектом разделились 50/50%. При этом все «инстанты» (референтные объекты) стоят в винительном падеже (примеры (4), (5)), все имена класса (нереферентные) – в родительном (примеры (6), (7)).

(4) *Разговор наш продолжался около часа, и я не знаю, до чего бы мы договорились, ежели б старик* ***не взял жену*** *за руку и со мною не распрощался.* [В.В. Измайлов. Ростовское озеро (1795)].

(5) *...Он* ***не любит свою жену****, несмотря на то, что она разумна, добронравна, домоводна, хороша и сама его любит.* [Н.И. Новиков. Живописец. Третье издание 1775 г. Часть I (1775)].

(6) ***Черемискїя жены не жертвуютъ*** *ни одному изъ сихъ боговъ, но для ихъ оставлены всѣ тѣ божества, которыхъ они въ женскомъ родѣ почитаютъ, и которыя суть нижеслѣдующїя.* [Н.П. Рычков. Журналъ или дневныя записки путешествїя Капитана Рычкова по разнымъ провинцїямъ Россїйскаго государства, въ 1769 и 1770 году (1770)].

(7) *Человек, который для показания остроты* ***не жалеет*** *матери,* ***жены****, сестер, братьев, друзей, каков бы умен ни был, достоин уничижения честных людей; ибо он нарушает добронравие и все должности.* [Екатерина II. [Полемика с Новиковым] (1769)].

С объектами *мать*, *дочь* и *невеста* вне зависимости от фактора референтности преобладает родительный падеж, примеры с винительным падежом единичны (по одному со словами *мать* и *невеста*). При этом слово *дочь* во всех примерах, и с родительным, и с винительным падежом, референтно, а доля родительного падежа у него больше (6 с родительным, 2 с винительным):

(8) *Некто из твоих подданных желает тебе благополучия и просит, чтобы ты* ***не отдавал твою дочь****, а нашу государыню Полкану, когда подступит он под город, а победить его принимает он это на себя; и ты увидишь в поле одного противника из твоих подданных столь многочисленному и свирепому воинству».* [М.Д. Чулков. Пересмешник, или Славенские сказки (1766-1768)].

(9) *Я* ***не вывозил*** *еще* ***дочери*** *на большие и нарядные съезды, но все любопытное показывал ей, не щадя денег.* [И.М. Долгоруков. Повесть о 6рождении моем, происхождении и всей моей жизни, писанная мной самим и начатая в Москве, 1788-го года в августе месяце, на 25-м году моей жизни / Часть 4 / 1799-1806 (1788-1822)].

Остальные имена класса (6 примеров) – в родительном падеже. Здесь зависимость выбора падежа от категории определенности/неопределенности соблюдена.

Единственный пример с именной группой – именем класса в винительном падеже –появляется в 1750-е гг. с объектом *мать* в цитировании евангелического текста:

(10) *Он же, не слушая матере, поступал по обычаям поганским, не веды бо закона божия, еже есть должно родителем повиноватися, ибо ежели не покоряется и благому наставлению не последует, той погибнет, яко ж речено есть: «Аще кто отца и* ***матерь не послушает****, смертию да умрет».* [В.Н. Татищев. История российская в семи томах. Том второй (1750)].

Таким образом, соответствие категории неопределенности соблюдено только в примерах с существительным *жена*. Остальные названия родства – и «инстанты», и имена класса – в большинстве сто-

[2] Адрес в Интернете: ruscorpora.ru

[3] С 1750-х гг., более ранних примеров с названиями родства основной корпус не дает.

ят в родительном падеже и от фактора определенности не зависят.

XIX век. За это столетие в основном корпусе найдено 336 примеров с названиями родства по женской линии (283 – с «инстантами» и 53 – с именами класса). Из них: 77 предложений со словом *мать* (а также 18 примеров с однокоренными словами: *мама, мамка, мамаша, матушка*), 49 – со словом *дочь*, 102 – *жена*, 24 – *сестра*, 66 – с другими названиями родства (*бабка, бабушка, внучка, крестная, кузина, кума, свекровь, сноха, невестка, падчерица, мачеха, тетя, тетушка, племянница, невеста*).

«Инстанты». Для всех названий родства в начале XIX в. преобладающим или даже единственным падежом является родительный (рис. 1), что демонстрирует почти неукоснительное следование действовавшей грамматическое норме. Ровно с середины века – в 1850-х гг. – доля винительного падежа начинает возрастать и в некоторые десятилетия (1860-е, 1870-е, 1890-е) даже превышает долю родительного падежа. К концу столетия доля винительного падежа на 17,4% превышает долю родительного падежа. Разрыв невелик, но в этот период намечается вариативность падежей, которая оформится в XX веке.

Исключение составляет слово *сестра*. Оно появляется как «инстант» в винительном падеже относительно поздно – только в 1854 г. (пример (11)). Но в этот период доля винительного падежа сразу сравнивается с долей родительного: 2 примера с родительным и 2 – с винительным падежами (примеры (11) и (12)). В этих примерах кроме определенности объекта для выбора винительного падежа есть и другие причины: в примере (11) – это инфинитив между глаголом с отрицанием и объектом, а также совершенный вид глагола, управляющего объектом; в примере (12) – это также совершенный вид и непрямой порядок слов с постановкой объекта в позицию темы.

(11) *Но мать Ашиля* ***не хотела уведомить сестру*** *о своем существовании письмом,*
все собиралась сама ехать с нею познакомиться… [Е.П. Ростопчина. Палаццо Форли (1854)].

(12) *И* ***родную сестру не пожалел*** *светлейший Меншиков.* [П.И. Мельников-Печерский. Старые годы (1857)].

Во второй половине XIX в. винительный падеж слова *сестра* уверенно доминирует над родительным: в 1860-х гг. – 66,67%, в 1890-х – 80% и в 1870-1880-х – 100%. Причиной такой «приверженности» винительному падежу может быть то, что в подавляющем большинстве случаев объект относится к конкретному человеку, т. е. он сильно референтен. То же можно сказать о слове *дочь* – оно почти не используется в качестве имени класса, но даже к концу века винительный падеж не превалирует, составляя 50% пропорцию с родительным (тем не менее, доминируя в 1860-1879-х гг. – 62,5% и 70% соответственно).

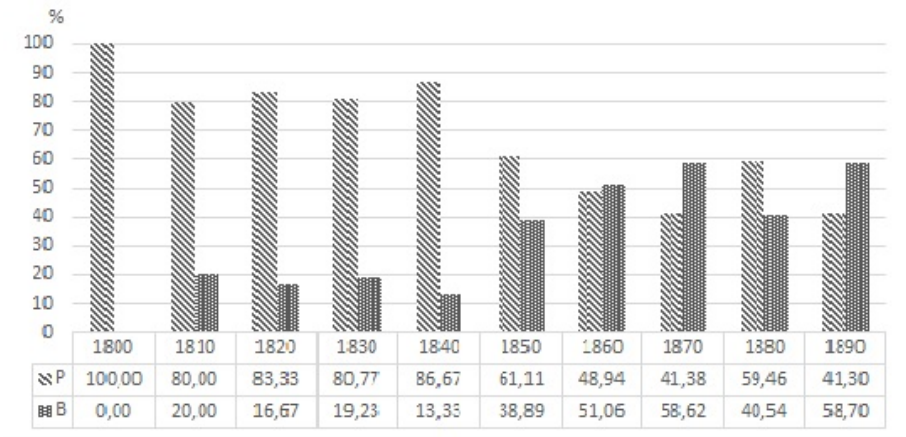
О действии фактора референтности в это столетие говорить не приходится: почти все примеры (за исключением примеров с глаголом *иметь*) содержат объект – название родства с референтным статусом, но большая их часть стоит в родительном падеже – как того требует грамматика.

Имена класса. В терминах родства – именах класса – винительный падеж ни в одно из десятилетий не затмевает родительный (рис. 2), только в 1890-х гг. число примеров с одним и другим падежами сравнивается, но не более того.

Регулярно винительный падеж в этой позиции начинает появляться, как и у «инстантов», с середины века с последующей – неравномерной – тенденцией к увеличению своей доли. Однако при сравнении графиков развития доли винительного падежа имен класса и «инстантов» (рис. 3) видно, что винительный падеж «инстантов» наращивает долю употребления более последовательно и к концу века устанавливает первенство над родительным падежом.

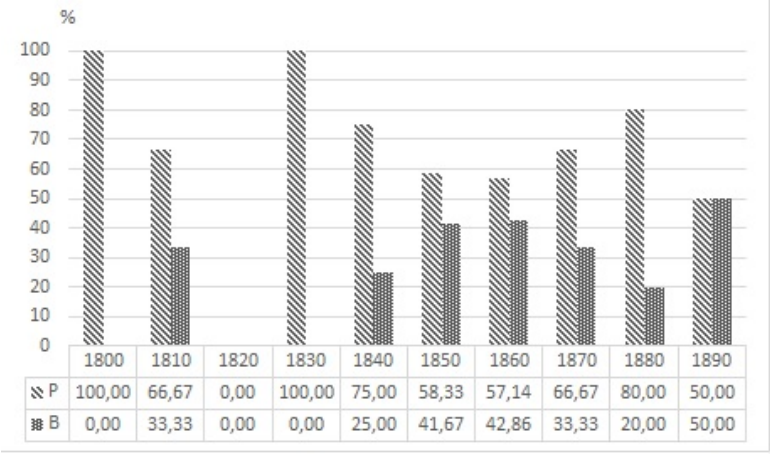
XX век. Основной корпус за XX в. представляет более объемную выборку: всего за столетие 812 примеров с названиями родства, из них 720 примеров – с «инстантами» и 92 – с именами класса.

«Инстанты». В двадцатом столетии в именах «инстантов» постоянно преобладает винительный падеж: 571 пример с винительным падежом и 149 – с родительным (рис. 4). Однако меняется величина разрыва между падежами: доля винительного падежа равномерно увеличивается. В начале века его доля ненамного превышала долю родительного падежа и составляла 56,36% (разница в 7 примеров), а в конце века его доля – 91,07%, и примеры с родительным падежом можно считать маргинальными (10 примеров – с родительным и 102 – с винительным).



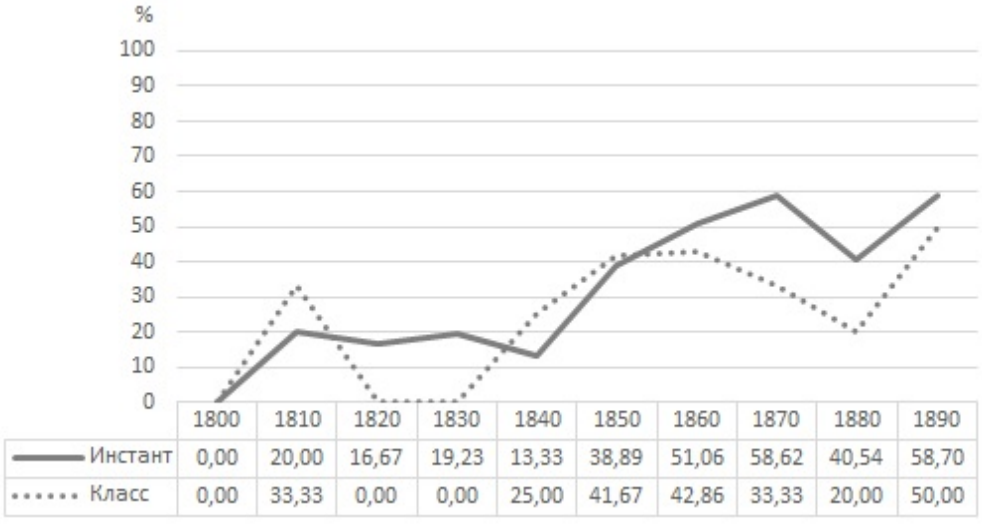
	1800	1810	1820	1830	1840	1850	1860	1870	1880	1890
Р	100,00	80,00	83,33	80,77	86,67	61,11	48,94	41,38	59,46	41,30
В	0,00	20,00	16,67	19,23	13,33	38,89	51,06	58,62	40,54	58,70

Рис. 1. Доля родительного и винительного падежей «инстантов» – терминов родства в XIX в. (%)



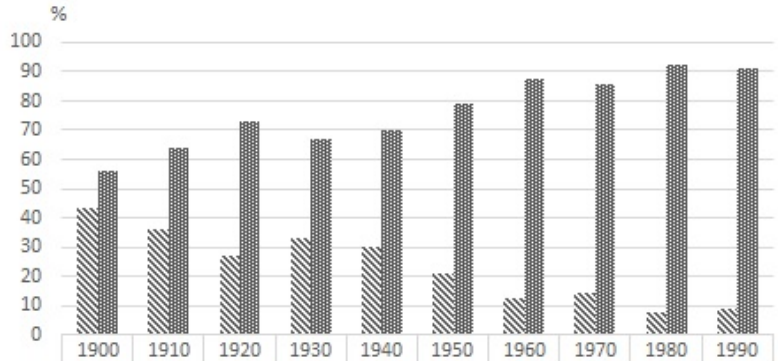
	1800	1810	1820	1830	1840	1850	1860	1870	1880	1890
Р	100,00	66,67	0,00	100,00	75,00	58,33	57,14	66,67	80,00	50,00
В	0,00	33,33	0,00	0,00	25,00	41,67	42,86	33,33	20,00	50,00

Рис. 2. Доля родительного и винительного падежей имен класса названий родства в XIX в. (%)



	1800	1810	1820	1830	1840	1850	1860	1870	1880	1890
Инстант	0,00	20,00	16,67	19,23	13,33	38,89	51,06	58,62	40,54	58,70
Класс	0,00	33,33	0,00	0,00	25,00	41,67	42,86	33,33	20,00	50,00

Рис. 3. Винительный падеж в названиях родства – имен класса и «инстантов» в XIX в. (%)



	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990
Р	43,64	35,90	26,76	32,93	30,23	20,69	12,50	14,44	7,78	8,93
В	56,36	64,10	73,24	67,07	69,77	79,31	87,50	85,56	92,22	91,07

Рис. 4. Доля родительного и винительного падежей «инстантов» названий родства в XX в. (%)

В этот период для «инстантов» отметим две закономерности.

Первая закономерность. В начале XX в. (1900-1910-е гг.) слова *мать, дочь* и *жена* (а это бо̀льшая часть примеров этого периода – 57 из 94) дают сравнительно низкие показатели винительного падежа. Так, в 1900-х и 1910-х у слов *мать* и *жена* доля падежей примерно равна, но в 1900-х несколько преобладает родительный падеж (доля винительного – 45,45% и 43,75% соответственно), а в 1910-х – чуть больше винительного (54,55% у слова *мать* и 57,14% у слова *жена*). Доля винительного падежа слова *дочь* в этот период и того меньше – 16,67% в 1900-е и 50% в 1910-е. При этом в следующее десятилетие, в 1920-е гг., доля винительного падежа везде возрастает: 56% – у слова *жена*, 75% – *мать*, 80% – *дочь*. Остальные названия родства уже в начале столетия дают более высокую долю винительного падежа – 63,64% в 1900-х и 62,5% в 1910-х, а слово *сестра* сразу «стартует» со 100%.

Эта особенность представляется продолжением тенденции XIX века: в 1890-х гг. у слова *дочь* доля винительного падежа составляла 50%, у слова *жена* – 47,37%, у слова *сестра*, напротив, – 80%. Сложнее объяснить её для слова *мать* – у него доля винительного падежа в конце XIX в. была 77,78% – и для остальных названий родства – у них было 33,33% винительного падежа.

Вторая закономерность. После всеобщего увеличения доли винительного падежа в 1920-е гг. в следующем десятилетии (в большинстве групп в 1930-е гг., у некоторых в 1940-е) отмечается его «проседание» до 50% (это заметно на рис. 4). Так, в группах с «инстантами» *мать, дочь, сестра* (в названный период эти группы представляют половину всех примеров – 40) в 1930-е гг. доля винительного падежа составила 50% (после 75%, 80%, 87% в 1920-е гг., соответственно). У существительного *жена*, а также менее частотных *бабушка, золовка, мачеха, невеста, племянница, тетя, тетушка* винительный падеж «проседает» чуть позже, в 1940-х гг. Объяснить эту закономерность пока затруднительно.

С середины века все группы существительных-«инстантов» показывают увеличение доли винительного падежа, что наблюдается у всех существительных рассматриваемой группы, кроме слов *мама, матушка, мамаша* – в этих словах – “неофициальных” соответствиях для слова *мать* – соотношение падежей не меняется: в начале столетия доля винительного падежа у них уже достаточно высока и во второй половине века продолжает оставаться на том же уровне.

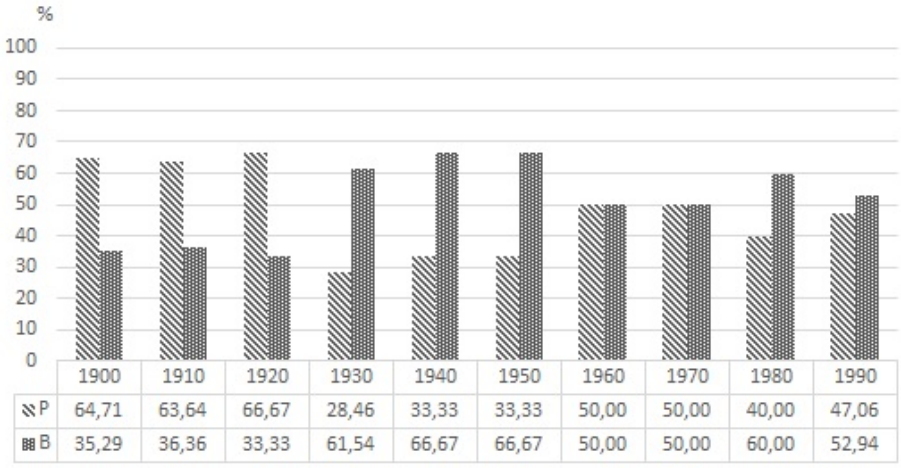
Имена класса. Примеров с родительным и винительным падежами в целом в именах класса поровну: 47 – с родительным и 45 – с винительным падежами (рис. 5).

В начале века доля родительного падежа больше, но уже с 1930-х гг. падежи «меняются местами»: винительный падеж резко начинает преобладать. В период 1960-1970-х гг. число родительного и винительного падежей сравнивается. Это равенство сохраняется и в конце века. Таким образом, с 1930-х гг. в падежном маркировании имен класса родительный падеж уступает место винительному падежу.

Наиболее закономерное развитие соотношения падежей – в группе со словом *мать* (рис. 6). В начале XX в. родительного больше (кроме периода 1920-1930-х гг.), а с 1960-х гг. происходит резкое увеличение доли винительного падежа (до 100%). В конце века примеров с обоими падежами почти поровну – разница в один пример (4 и 5).

Изменение доли винительного падежа имен класса и «инстантов» отражено на рис. 7.

На протяжении всего XX в. увеличивается доля винительного падежа обеих категорий. У «инстантов» эта доля растет более равномерно, чем у имен класса, и дает больший интервал (разница между 1990-ми и 1900-ми составляет 34,71% для «инстантов» и 17,65% для имен класса от общего количества примеров).



	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990
Р	64,71	63,64	66,67	28,46	33,33	33,33	50,00	50,00	40,00	47,06
В	35,29	36,36	33,33	61,54	66,67	66,67	50,00	50,00	60,00	52,94

Рис. 5. Доля родительного и винительного падежей имен класса названий родства в XX в. (%)

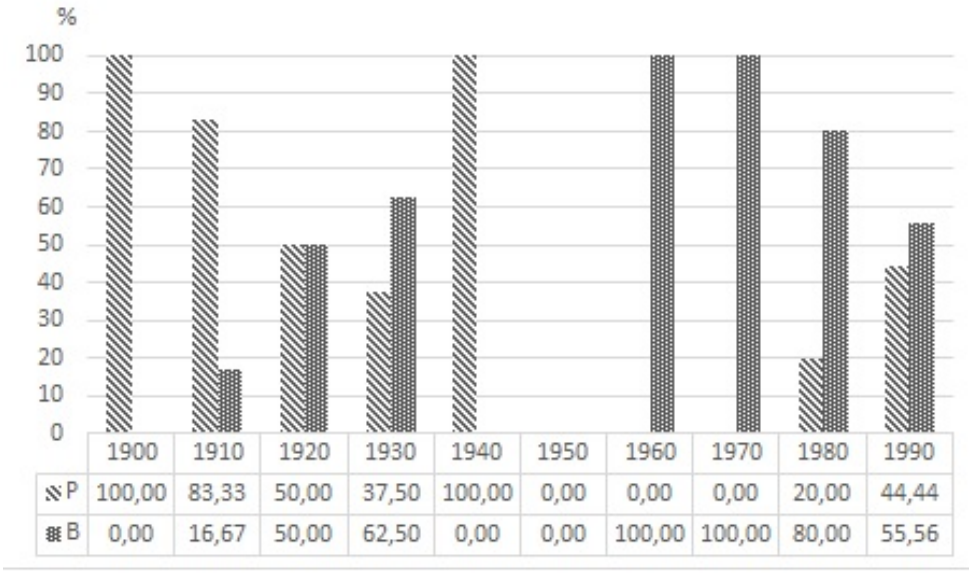


Рис. 6. Доля родительного и винительного падежей имен класса слова *мать* в XX в. (%)

Рис. 7. Доля винительного падежа названий родства – имен класса и «инстантов» в XX в. (%)

О действии фактора определенности можно было бы говорить в отношении «инстантов»: они в основном стоят в винительном падеже. Однако почти равное число примеров с обоими падежами имен класса не дает возможности сделать вывод о последовательном действии фактора определенности.

XXI век. Всего за 17 лет XXI в. в основном корпусе найден 221 пример: 201 – с «инстантами» и 20 – с именами класса. В целом XXI в. демонстрирует преобладание примеров с винительным падежом объектов – названий родства.

«Инстанты». В оба десятилетия XXI в. винительный падеж значительно преобладает (рис. 8).

Во всех группах существительных доля винительного падежа составляет от 87,5% до 100%. Так, абсолютное доминирование винительного в оба десятилетия показывают названия родства *дочь, мама, матушка, мамаша, теща, свекровь, невестка, сноха, племянница, внучка, бабушка, бабка*. Существительное *жена* дает 100% винительный падеж только в 2010-х гг.

Имена класса. В период с 2000 по 2017 гг. имена класса представляют всего 20 примеров: первое десятилетие XXI в. отмечено преобладанием винительного падежа (73,68% – 14 примеров); во втором десятилетии – только один пример, поэтому 100% винительного падежа в графике рис. 9 – достаточно условны. Тем не менее даже эти данные дают представление о тенденции падежного маркирования названий родства – имен класса в XXI в. Так, видно, что родительный падеж для этой группы больше не является основным – продолжается тенденция, намеченная в XX в.

Таким образом, к XXI в. имена класса и «инстанты» сходятся в общей тенденции к преобладанию винительного падежа для названий родства. Исходя из рассматриваемых примеров, условиями для появления родительного падежа можно считать усиление отрицания с помощью двойной частицы *ни… ни…* или глагола *иметь*. Однако, как показывают примеры, даже при этих условиях винительный падеж возможен:

(13) *И Нюша понимала: никогда, никогда, никогда* ***не увидит ни мать, ни дочь***. [Галина Шергова. …Об известных всем (2002-2004)]

(14) *Конечно, провожая Ирину в Москву, но, кто не рискует, тот* ***не имеет такую жену***. [Эльвира Савкина. «Я похожа на маленькую симпатичную обезьянку» (2002) // «Дело» (Самара), 26.04.2002]

Ожидаемой закономерности («инстанты» в винительном падеже, имена класса – в родительном) нет.

Изменение доли винительного падежа с XVIII до XXI вв.

Графики, отражающие изменение доли родительного и винительного падежей с 1750 по 2017 гг., представим на рис. 10-12.

«Инстанты». В развитии падежного соотношения названий родства – «инстантов» выделяются периоды (рис. 10).

1750-1790 гг. – попеременное лидерство.

В XVIII в. в примерах с названиями родственниц попеременно преобладает то родительный, то винительный падеж. В 1780-х гг. падежи разделились поровну. Этот период можно назвать периодом попеременного лидерства.

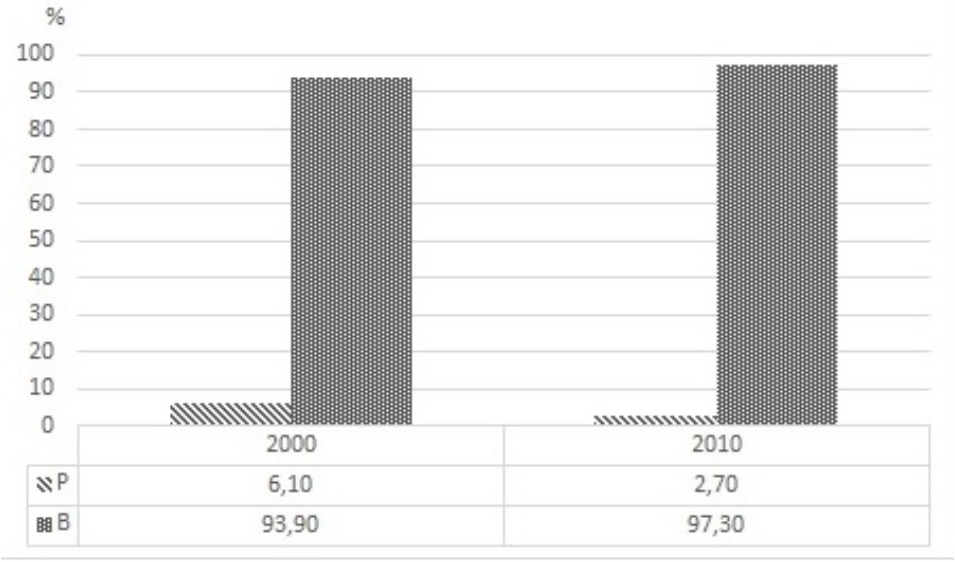


Рис. 8. Распределение родительного и винитнльного падежей названий родства – «инстантов» в XXI в. (%)

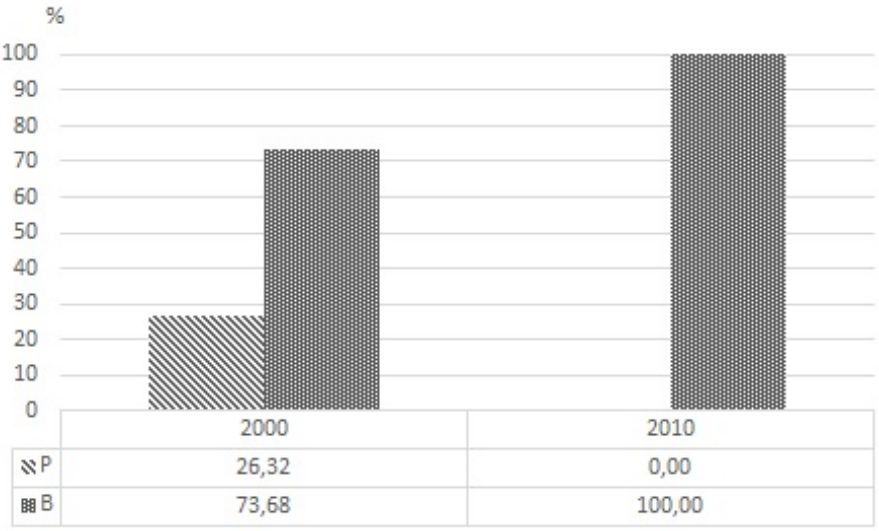


Рис. 9. Распределение родительного и винительного падежей названий родства – имен класса в XXI в. (%)

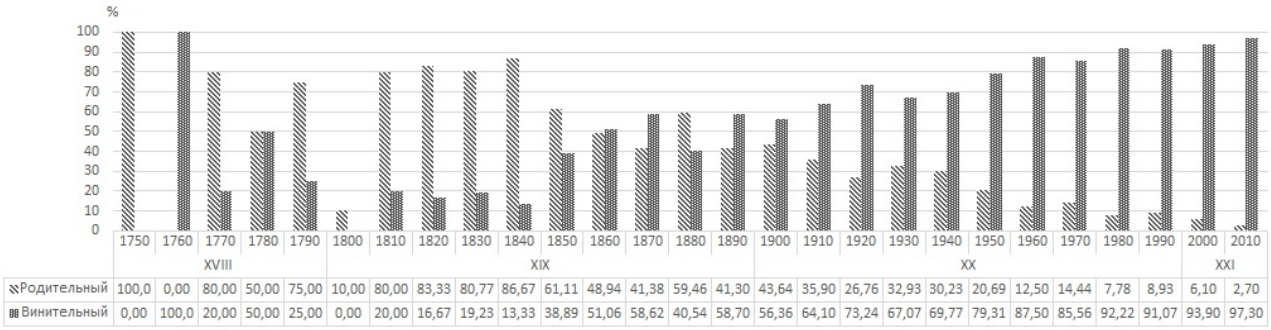


Рис 10. Распределение доли родительного и винительного падежей «инстантов» названий родства в период 1750-2017 гг. (%)

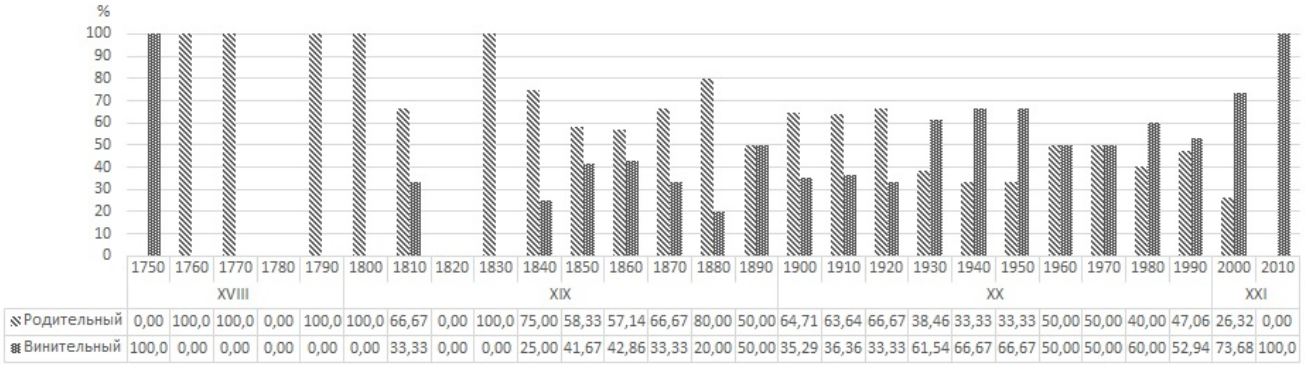


Рис. 11. Распределение доли родительного и винительного падежей имен класса – названий родства в период 1750-2017 гг. (%)

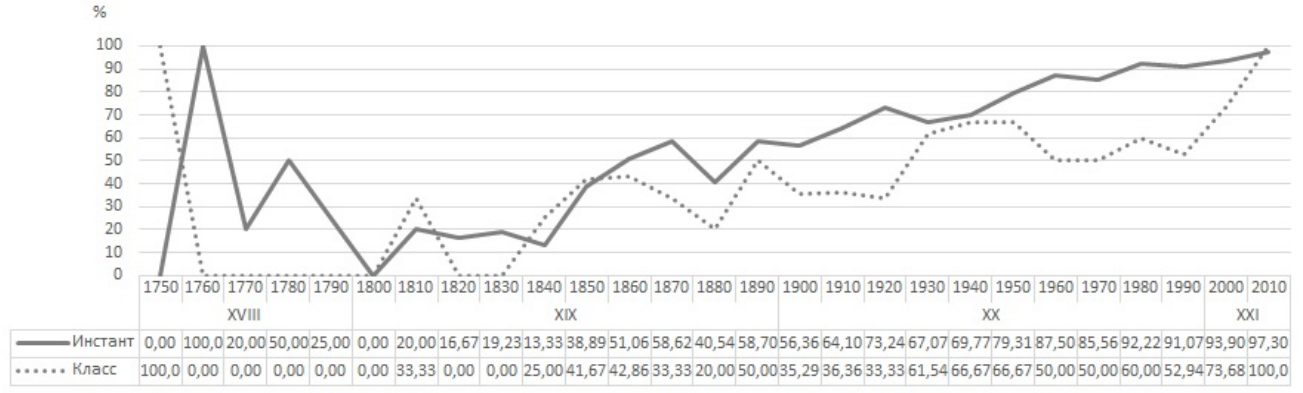


Рис. 12. Изменение доли винительного падежа имен класса и «инстантов» названий родства в период 1750-2017 гг. (%)

1790-1840 гг. – период родительного падежа.

С 1790-х гг. становится больше родительного падежа. С начала XIX в. родительный падеж устанавливает очевидное доминирование.

1850-1900 гг. – «сближение» падежей.

С 1850-х гг. доли падежей сближаются и разница между родительным и винительным составляет 22,22% (ср. с разрывом в 73,34% в 1840-х). Далее эта разница колеблется между 2,12% (в 1860-х) и 18,92% (в 1880-х). Этот период сближения продолжается до 1900-х гг.

1910-2010 гг. – период роста винительного падежа.

С 1910-х гг. винительного падежа становится больше. С этого момента его доля растет – до 97,3% в 2010-х.

Имена класса. Соотношение падежей в терминах родства – именах класса также делится на периоды (рис. 11):

1760-1830 гг. – период монополии родительного падежа.

В 1750-х гг. НКРЯ дает единственный пример с винительным падежом (пример (10)). Он может быть отражением более ранней тенденции к доминированию винительного падежа. Однако сделать такой вывод на основании одного примера невозможно, и в рамках данного исследования его следует признать маргинальным и не включать в выводы.

С 1760-х и до 1830-х гг. включительно имена класса представлены только в родительном падеже (кроме 1810-х гг., где родительный все равно считается доминирующим).

1840-1920 гг. – начало конкуренции при доминировании родительного падежа.

Винительный падеж появляется в 1840-х гг., но до 1920-х гг. его доля не превышает долю родительного падежа, который все еще остается главным в этот период, но это уже отношения не монополии, а конкуренции.

1930-2010 гг. – винительный падеж выходит вперед.

С 1930-х гг. винительного падежа становится больше, его доля превышает долю родительного и постепенно увеличивается, а в XXI в. уже больше винительного падежа.

Как видно, приведенные периоды у «инстантов» и имен класса не совпадают полностью. Однако можно выделить точки на временно̀й шкале, когда в обеих группах происходили изменения тренда:

- до 1830-х гг. в обеих группах преобладает родительный падеж. В именах класса это доминирование изначально, в «инстантах» оно появляется с конца XVIII в.;
- 1830-1840-е гг. – выходит вперед винительный падеж, и между двумя падежами начинается конкуренция;
- 1910-1930-е гг. – в падежном соперничестве начинает побеждать винительный падеж; с этого периода его доля неуклонно растет в обеих группах.

Таким образом, развитие падежного соотношения в именах класса и «инстантах» следует единой тенденции (рис. 12), но есть и различия:

1) во второй половине XVIII в. имена класса не стоят в винительном падеже (за исключением 1750-х гг. с одним примером). В «инстантах» в это же столетие винительный падеж не исключен;

2) с середины XIX в. (с 1850-х гг.) в обеих группах увеличивается количество примеров с винительным падежом. У «инстантов» их рост последовательный и равномерный; у имен класса заметны подъемы (1930-1950) и «провалы» (1900-1920; 1960-1990) в этом движении. В целом доля винительного имен класса ниже, чем «инстантов». К концу наблюдаемого периода (в 2000-е гг., поскольку имена класса в 2010-е представляют только 1 пример) имена класса приближаются в этом движении к «инстантам», но не сравниваются с ними.

Изменение доли винительного падежа у имен существительных

Необходимо отметить, что разные группы названий родства демонстрируют неодинаковую динамику доли винительного падежа «инстантов». Можно выделить две тенденции: 1) у существительных *мама, матушка, мамаша* и *сестра* и 2) у остальных существительных – названий родства. Эти тенденции проиллюстрируем графиками, отражающими долю винительного падежа группы «инстантов» (рис. 13, 14). Шаг изменения – 10 лет.

1. Слова *мама, матушка, мамаша, сестра*

В отличие от других исследуемых названий родства слова *матушка, мама, мамаша* характерны для разговорного или эмоционально-пренебрежительного контекста. Можно считать, что эти слова больше тяготеют к референтности: носители языка эмоционально окрашивают высказывания о денотатах, с которыми хорошо знакомы, т.е. с актуализованными объектами.

Эмоционально-стилистическая окрашенность этих слов может быть причиной того, что изначально (с XIX в., когда примеры с указанными словами начинают фиксироваться в НКРЯ) доля винительного падежа достаточно высока (хотя родительного падежа в этот период все же больше). Уже в следующий период – во вторую половину XIX в. – доля винительного падежа постоянно значительно выше доли родительного.

Схожее развитие доли винительного падежа демонстрирует существительное *сестра* (рис. 13). Примеров со словом *сестра* в целом по основному корпусу НКРЯ немного. Первые примеры появляются только в начале XIX в., причем с полным преобладанием родительного падежа. Как и у эмоциональных слов *мама, матушка, мамаша,* со второй половины XIX в. (1870-х гг.) доля винительного падежа приближается к 100% и на протяжении XX в. удерживает эти позиции. Исключение составляют 1930-1940-е гг. для названия родства *сестра*, когда родительный падеж вновь вступил в борьбу, и некоторые колебания в этой же группе во второй половине XX в.

Причина того, что слова родства *мать, матушка, мамаша* тяготеют к винительному падежу, ясна – она заключается в их эмоционально-оценочной окрашенности. По этой же причине их изменение в XX-XXI вв. так последовательно и однозначно. Вопрос, однако, в том, почему существительное *сестра* следует этой же тенденции, пусть и с некоторыми отступлениями. Вряд ли этот термин родства более тяготеет к актуализованным денотатам, чем, скажем, *дочь*. В нашем исследовании этот вопрос приходится оставить открытым.

2. Слова *мать, дочь, жена* и другие

Динамика винительного падежа существительных *мать, дочь* и *жена* несколько отличается. У них, как видно на рис. 14, распределение доли винительного падежа в XVIII-XXI вв. следует общей тенденции.

В XVIII в. доля винительного падежа в разные десятилетия то поднимается до предела, то опускается до 0. Необходимо учитывать, что в каждой группе в этот период малое число примеров, поэтому выводы о нём неточны.

На границе XVIII-XIX вв. (1790-1800) число примеров с винительным падежом сильно уменьшается по сравнению с родительным; с 1820-1840-х гг. есть небольшое число примеров с винительным падежом,

но всю первую половину XIX в. родительный падеж остается доминирующим.

С середины XIX в. (1850-1860) винительный падеж вступает в конкуренцию с родительным падежом и начинает в ней побеждать. Эта борьба продолжается весь XX в. с переменным успехом. Во второй половине XX столетия (1950-1960) винительный падеж выходит вперед и до конца рассматриваемого нами периода сохраняет эти позиции, почти полностью вытесняя родительный падеж в терминах родства.

Остальные названия родства (*тетя, бабушка, племянница, свекровь, невестка, сноха* и пр.) в полной мере можно отслеживать с первой половины XIX в. Однако можно предположить, что они следуют той же тенденции, что и слова *мать, дочь, жена.*

Несмотря на то, что для групп названий родства кривые получились неодинаковые, заметна общая тенденция (рис. 15):

- в конце XVIII – первой половине XIX вв. у всех групп доля винительного падежа очень низка;
- во второй половине XIX в. у всех групп наблюдается резкое увеличение доли винительного падежа;
- начиная со второй половины XIX в. у всех групп с разной динамикой, но одинаково неуклонно увеличивается доля винительного падежа;
- как результат – в XXI в. во всех группах винительный падеж уверенно доминирует над родительным.

Эта общая тенденция отражена в графике на рис. 15.

Подобные нашим результаты были получены М.В. Русаковой [10], работавшей с тремя выборками с непрезумптивной лексикой из НКРЯ основного корпуса XVIII-XXI вв. В её исследовании графическое выражение динамики изменения в основной выборке сходно с полученным в нашем исследовании. Разница – в скорости нарастания доли винительного падежа в XX в.: для презумптивных имен – названий родства – увеличение доли винительного падежа в этот период происходит более резко, чем в общем исследовании М.В. Русаковой: в ее исследовании в начале века 10% примеров с винительным падежом, в конце века – 40% [10, с. 332]; у нас для презумптивной лексики «инстантов» это 56% и 91%, соответственно. Более резкое увеличение доли винительного падежа презумптивных имен может объясняться их константно референтным статусом. Для подтверждения этой идеи необходимо подобное изучение других групп презумптивных имен.

М.В. Русакова приводит убедительное объяснение наблюдаемой тенденции.

Низкий процент винительного падежа на рубеже XVIII-XIX вв. объясняется стремлением следовать грамматической норме, диктуемой в грамматике Н.И. Греча [6]: «Если даже Пушкин воспринимает грамматику как «руководство к действию», то для других носителей языка она, скорее всего, была просто догмой» [10, с. 349].

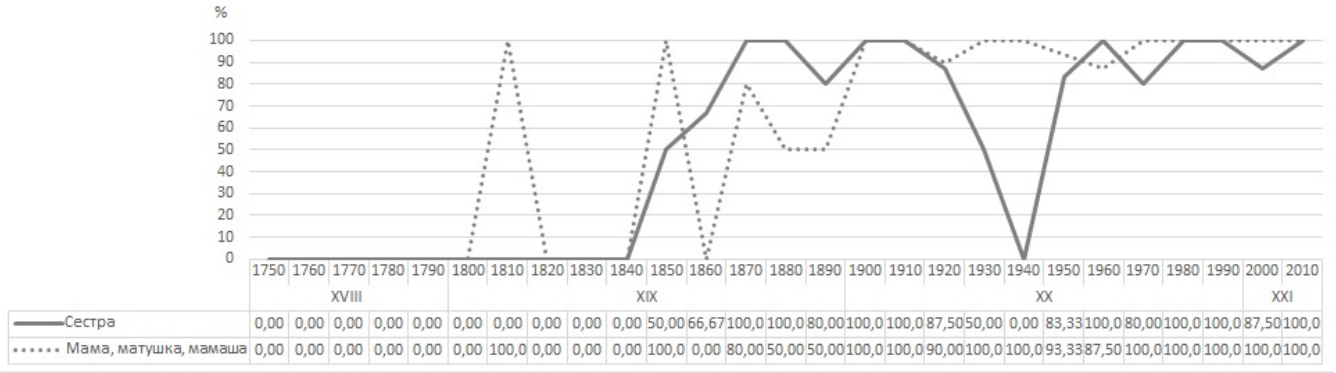


Рис. 13. Изменение доли винительного падежа «инстантов» *мать, матушка, мамаша* и *сестра* в XVIII-XXI вв. (%)

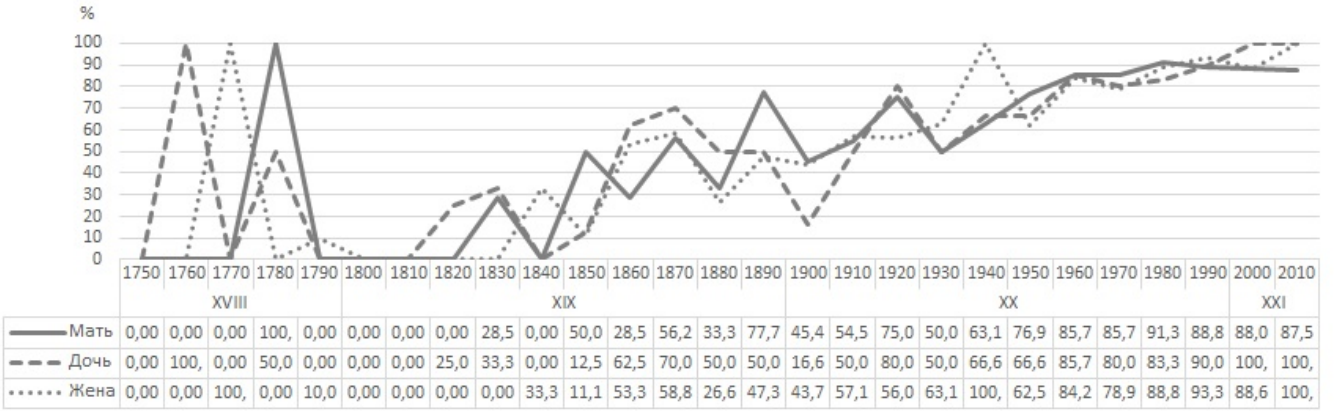


Рис. 14. Изменение винительного падежа «инстантов» *мать, дочь, жена* в XVIII-XXI вв. (%)

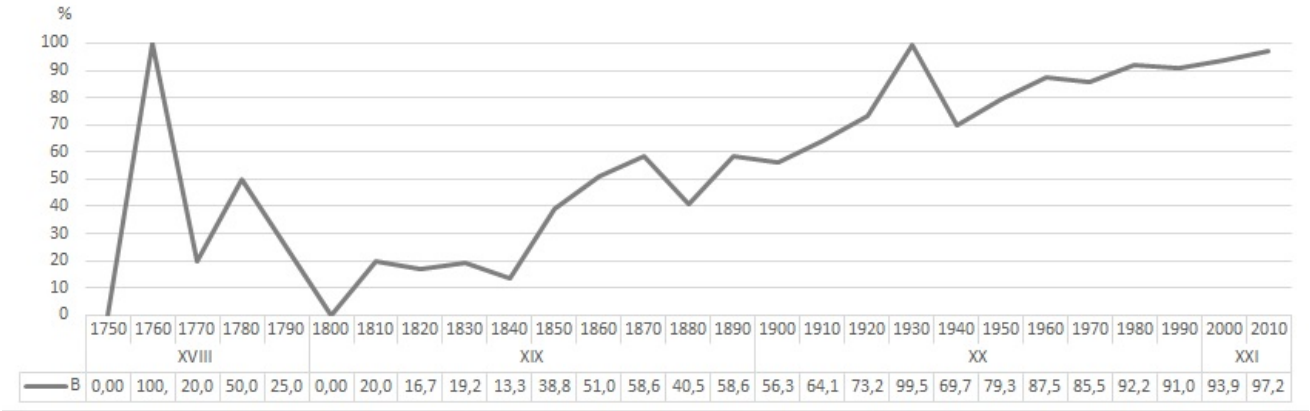


Рис. 15. Средние показатели изменения доли винительного падежа названий родства в XVIII-XXI вв. (%)

Увеличение доли винительного падежа в середине XIX в. объясняется отходом от грамматической нормы пишущих людей, в большинстве своем объединенных демократическими настроениями: «Во второй половине XIX века, в разночинный период русской словесности, круг литераторов, и вообще пишущих людей, значительно расширился, престиж литературной нормы у демократически, а то и революционно настроенной пишущей публики, соответственно, понизился – видимо, это и привело к первой волне нарастания аккузативного оформления прямого объекта при отрицании» [10, с. 350].

Нарастание доли винительного падежа в середине XX в. М.В. Русакова объясняет смешением форм родительного и винительного падежей: «Думается, что в значительном количестве случаев, употребляя форму генитива при отрицании, носители русского языка воспринимают ее и «немножко» как форму аккузатива» [10, с. 352]. Это объяснение представляется спорным; более вероятно влияние других факторов (см., например, списки факторов в [3, 5]), не связанных с неумением носителей языка отличить одну падежную форму от другой. В нашем исследовании не выявлено резкого повышения доли винительного падежа в середине XX в., наблюдаемого М.В. Русаковой, поэтому позволим себе оставить поиск причин для дальнейших исследований.

ЗАКЛЮЧЕНИЕ

Проведенное нами корпусное исследование показало, что с XVIII в. фактор определенности не имел постоянного влияния на падежное маркирование объекта при отрицании. В противном случае рассмотренные примеры показывали бы соответствие между соотношением падежей в двух группах названий родства и определенностью объекта.

Выбор родительного или винительного падежа в отрицательной конструкции в XVIII-XIX вв. зависел скорее не от внутренней имплицитной семантики предложения, а от внешних факторов, а именно: ориентируется пишущий на требования грамматики, опирающейся исключительно на синтаксическую конструкцию, или нет. Этот вывод подтверждается динамикой соотношения падежей в каждый из рассмотренных периодов; к таким же выводам приходит М.В. Русакова, работавшая с непрезумптивным материалом.

До XIX в. винительный падеж почти не уступал родительному по числу употреблений. С 1800-х гг. наблюдается резкое увеличение доли родительного падежа объектов – названий родства, чему сопутствует рост влияния грамматики, в частности Н.И. Греча. С 1850-х гг., с приходом в литературу разночинцев, число примеров с винительным падежом увеличивается. Начиная с XX в., когда число неграмотных носителей языка резко снижается, названия родства маркируются в равной мере обоими падежами, а во второй половине столетия, когда в школьных учебниках не указывается требование замены падежа объекта при отрицании, а на академическую грамматику ориентируются немногие носители языка, родительный падеж уступил позиции винительному. По-видимому, в этот же период вступают в действие семантические факторы, многократно описанные в научной литературе. Возникшая в XX в. тенденция продолжилась и в новом тысячелетии. Таким образом, картина второй половины XX в. – начала XXI в. сходна с соотношением падежей в середине XVIII в.

Сравнение с результатами исследования М.В. Русаковой приводит к еще одному наблюдению: в XIX в. в презумптивных именах увеличение доли винительного падежа произошло более динамично, чем в непрезумптивных, а в XXI в. презумптивные имена дают более высокий процент винительного падежа (для некоторых названий родства – 100%), чем непрезумптивные. Следовательно, некоторое влияние категория определенности оказывала, пусть не настолько явное, как ожидалось.

Основной корпус НКРЯ дает слишком мало примеров XVIII в., чтобы можно было говорить о закономерности, а в XIX в. пишущие носители языка уже следовали требованиям грамматики. Более определенные результаты может дать исторический анализ текстов более раннего периода, предшествующего созданию русского литературного языка. Тогда можно будет предположить, что носители языка, отсту-

пив от требований грамматики, не расшатали норму, как считалось, а вернулись к «дограмматическому» способу выбора падежа. Чтобы это можно было утверждать с уверенностью, потребуются новые исследования более широкого круга презумптивной лексики в отрицательной конструкции.

СПИСОК ЛИТЕРАТУРЫ

1. Томсон А.И. Винительный падеж прямого дополнения в отрицательных предложениях в русском языке. Отд. отт. из «Русского Филологического Вестника». – Варшава: Типография Варшавского Учебного Округа, 1902. – 43 с.
2. Ицкович В.А. Очерки синтаксической нормы // Синтаксис и норма / отв. ред. Г.А. Золотова. – М.: Наука, 1974. – С. 43-106.
3. Mustajoki A., Hannes Heino. Case selection for the direct object in Russian negative clauses. Part II: Report on a Statistical Analysis: Slavica Helsingiensia 9. – Helsinki: Department of Slavonic Languages, University of Helsinki, 1991.
4. Падучева Е.В. Генитив дополнения в отрицательном предложении // Вопросы языкознания. – 2006. – № 6. – С. 21–43.
5. Русская грамматика: в 2 т. / Академия наук СССР, Институт русского языка. – Т. 2. Синтаксис. – М.: Наука, 1980. – 710 с.
6. Практическая русская грамматика, изданная Николаемъ Гречемъ. – Санктпетербургъ: Типографія Императорскаго Санктпетербургскаго Воспитательнаго Дома, 1827. – 588 с.
7. Крылов С.А. Детерминация имени в русском языке: теоретические проблемы // Семиотика и информатика. – 1997. – Вып. 35. – С. 244–271.
8. Булыгина Т.В., Шмелев А.Д. Пространственно-временная локализация как суперкатегория предложения // Вопросы языкознания. –1989. – № 3. – С. 51-61.
9. Шмелев А.Д. Русский язык и внеязыковая действительность. – М.: Языки славянской культуры, 2002. – 492 с. – (Сер. «Язык. Семиотика. Культура»).
10. Русакова М.В. Аккузатив *vs*. генитив в конструкциях с отрицанием: динамика взаимоотношений на историческом отрезке, отраженном текстами в НКРЯ // Элементы антропоцентрической грамматики русского языка. – М.: Языки славянской культуры, 2013. – С. 324–353.

Материал поступил в редакцию 25.03.19.

Сведения об авторе

ФЕСЕНКО Вера Павловна – преподаватель Высшей школы экономики, Москва
e-mail: verun4ik_18@mail.ru

УДК 81’322.2:821.161.1

С.Н. Полохин, Б.Я. Рябко, Н.Н. Савина

Развитие теоретико-информационного метода классификации текстов и его применение к анализу авторства*

Теоретико-информационный подход, базирующийся на применении методов сжатия данных, применен к анализу романов «12 стульев» и «Золотой теленок» для установления авторства. Проведенное исследование позволяет сделать вывод о том, что авторами являются И. Ильф и Е. Петров. Однако при этом обнаруживается некоторое сходство с произведениями М.Булгакова и К.Паустовского.

Ключевые слова: *атрибуция, теоретико-информационные методы атрибуции, сжатие данных, русская литература XX в.*

ВВЕДЕНИЕ

Атрибуция в литературоведении (от лат. *attribution* – приписывание) – это установление автора художественного произведения. Существует не одно литературное произведение, авторство которого когда-то было или до сих пор является неустановленным, ложным или сомнительным. Возможно, самая известная проблема атрибуции – так называемый «Шекспировский вопрос». Многие исследователи сомневаются в авторстве Шекспира в отношении целого ряда произведений [1]. В последние десятилетия при решении проблемы атрибуции литературных произведений наряду с традиционными литературоведческими методами стали применять подходы таких дисциплин, как математическая статистика, теория случайных процессов, распознавание образов и ряда других научных дисциплин [2, 3].

Проблема атрибуции известна и для произведений на русском языке [4]. Среди многочисленных примеров отметим только, что произведение «Роман с кокаином», опубликованное в 1934 г. под именем М. Агеева, как позже выяснилось, было написано Марком Леви [5], а фельетон «Главполитбогослужение», публиковавшийся в сборнике рассказов В.П. Катаева, как позднее утверждал сам Катаев, был написан М.А. Булгаковым.

В 2013 г. Ириной Амлински опубликована книга «12 стульев от Михаила Булгакова», в которой приводились рассуждения и доводы в пользу того, что романы «12 стульев» и «Золотой теленок» были написаны не И. Ильфом и Е. Петровым, а М. Булгаковым [6]. После этого мнения литературоведов, писателей и читателей разделились. Есть те, кто уверен, что мистификация была [7], те, кто уверен, что ее не было [8], и наконец те, кто так и не смог прийти к окончательному выводу [9].

В существующих работах об авторстве произведений «12 стульев» и «Золотой теленок» все исследования литературоведческие. Производится анализ стиля авторов, берутся во внимание всевозможные отсылки к произведениям М. Булгакова, рассматриваются исторические аспекты написания произведений. Несмотря на весь объем проделанных исследований, специалисты не пришли к единому выводу. Цель нашей работы – исследовать вопрос авторства этих романов, прибегая к теоретико- информационному подходу, основанному на использовании методов сжатия данных.

Важно отметить, что существует не один подход к решению задачи атрибуции, использующий математические методы. Среди известных отметим методы, построенные на применении авторского инварианта [10] или модели цепей А.А. Маркова [11, 12], подходы, базирующиеся на векторизации текста в совокупности с методами машинного обучения [13, 14]. В работах [15, 16] для решения задачи атрибуции используется теоретико-информационный подход, основанный на сжатии данных [17], который доказал свою эффективность в сравнении, например, с подходом, применяющим Марковские цепи [12].

На основе этого подхода получен метод, позволяющий получать статистически обоснованные результаты. При его применении к проблеме атрибуции текстов «12 стульев» и «Золотой теленок» показано, что наиболее вероятными авторами являются И. Ильф и Е. Петров, однако полученные результаты демонстрируют, что влияние Булгакова на эти произведение довольно существенное.

ОПИСАНИЕ МЕТОДА

Сначала кратко опишем схему применения теоретико-информационного подхода, а затем экспериментально найдем значения параметров, повышающих его эффективность, используя представительный набор прозаических произведений писателей – современников Ильфа, Петрова и Булгакова.

* Работа выполнена при поддержке Российского фонда фундаментальных исследований (грант № 18-29-03005).

Алгоритмы сжатия данных

В настоящее время во многих информационных технологиях используются методы сжатия данных, реализованные в виде компьютерных программ-архиваторов. На их вход подаются текстовые данные, которые они кодируют в файлы меньшей длины, т.е. «сжимают». Это достигается за счет того, что архиваторы находят неравномерность частоты встречаемости символов и их комбинаций, используя для этого различные методы теории информации и других дисциплин, включая описание сжимаемого текста при помощи формальных грамматик, а также другие подходы, присущие искусственному интеллекту.

Схема использования методов сжатия для классификации текстов может быть показана на схематическом примере: даны три текста *A, B, C*, причем известно, что *A* и *B* порождены разными источниками информации, а о третьем известно, что он порожден одним из двух, но неизвестно каким. Это – задача классификации или «проблема трех выборок» в математической статистике. При этом источники информации – авторы, известные для текстов *A* и *B*, тогда как авторство *C* точно неизвестно.

Суть метода – объединеняем тексты в пары *AC* и *BC* и «сжимаем» первую и вторую, затем вычисляем разности {*AC*} – {*A*} и {*BC*} – {*B*}, где фигурные скобки обозначают длину сжатого файла . Если {*AC*} – {*A*} меньше {*BC*} – {*B*}, то делаем вывод, что *A* и *C* порождены одним источником. Если меньше {*BC*} – {*B*}, то один и тот же автор у этой пары. Вывод основывается на том, что архиватор, сжимая вторую часть (*C*), использует статистические особенности, найденные им при сжатии первой части (*A* или *B*), и, следовательно, текст *C* сжимается лучше после текста, порожденного «его» источником.

Опишем этот подход более формально для случая нескольких текстов, написанных разными авторами. Пусть X_i – некоторый текст автора – длина текста *Y*. Пусть *XY* – текст, к которому справа без каких-либо дополнительных символов приписали текст *Y*. Наконец, пусть Z_c – сжатый некоторым архиватором текст *Z*.

Величина $CCC\left(Y\middle|X_i\right)=\left|\left(X_iY\right)_c\right|-\left|X_{i_c}\right|$, называемая условной степенью сжатия текста, характеризует степень близости X_i и *Y*. Неформально предполагается, что чем лучше был сжат текст X_iY, тем больше информации об *Y* содержалось в X_i, а это означает, что стиль *i*-го автора в наибольшей мере совпадает со стилем, в котором написан текст *Y* неизвестного автора.

Схема применяемого алгоритма

При использовании теоретико-информационного подхода многие исследователи рекомендуют разбивать сравнительно большие тексты на части, ориентировочно равные нескольким страницам, что позволяет применять статистичекие методы для повышения надежности выводов. Способ разбиения на части и размер этих частей могут влиять на эффективность метода

Опишем подход формально. Пусть имеется *N* авторов, *i*-му автору принадлежит некоторый текст X_i, i = 1,2, …, *N*. Пусть имеется текст *Y* неизвестного автора и для него требуется определить среди *N* имеющихся авторов наиболее вероятного. Для этого разобьем каждый текст X_i на блоки $\left\{X_{ij}\middle|j=1..M_i\right\}$ одинакового размера S_1, то есть $\left|X_{ij}\right|=S_1$, $i=1..N$, $j=1..M_i$ (везде далее за $\left|X\right|$ будем обозначать длину блока *X*). Аналогично, разобьем на блоки размера S_2 текст неизвестного автора *Y*. Для каждого i-го автора зафиксируем некоторое подмножество I_i^{train} множества номеров его блоков, $\left|I_i^{train}\right|=K_1$. Для *i*-го автора назовем его блоки с номерами I_i^{train} обучающей выборкой – с ними будут сжиматься блоки текста неизвестного автора. Обозначим $Q_i^{train}=\left(\sum_{j\in I_i^{train}}X_{ij}\right)$, где под суммой текстов понимается их конкатенация (соединение). Введем для каждого автора счетчик количества случаев, когда исследуемый фрагмент лучше сжимается при добавлении к его обучающей выборке текстов и положим начальное значение этого счетчика равным нулю. Каждый блок текста *Y* сожмем с текстом обучающей выборки каждого автора. Для автора, с текстами которого блок сжался лучше всего, увеличим значение счетчика на единицу. Тогда наиболее вероятный автор *A* текста *Y* – тот, чей счетчик принимает максимальное значение. Формально алгоритм можно записать в виде:

$$a=\left(\underbrace{0,0,...,0}_{N}\right)-\textit{счетчики}$$

Для $k=1..K_1$

  Для $i=1..N$

$$C_i=\left|Q_i^{train}Y_k\right|-\left|Q_i^{train}\right|$$

$$a_{\arg\min(\{C\})}+=1$$

$$A=\arg\max(a)-\textit{автор исследуемого текста}$$

Пусть, например, имеются тексты произведений «Мои университеты» Максима Горького и «В дни поражений и побед» Аркадия Гайдара. Определим с помощью алгоритма, кто из этих двух авторов является автором произведения «Вечер у Шамова». В соответствии с алгоритмом, разобьем тексты «Мои университеты» и «В дни поражений и побед» на непересекающиеся блоки одинакового размера, а затем выберем среди блоков каждого произведения, например, по 10 блоков и объединим для каждого автора его блоки в единый текст в случайном порядке. Теперь разобьем на блоки текст произведения «Вечер у Шамова». Пусть их будет 12. Каждый из этих блоков попробуем сжать с ранее полученными текстами. Если блок лучше сжался с текстом М. Горького, то к «его» счетчику прибавляем 1, иначе 1 прибавляем к «счетчику» А. Гайдара. Наибольшее значение получилось у счетчика Максима Горького, и именно он на самом деле является автором рассказа «Вечер у Шамова».

Как видно, алгоритм обладает рядом настраиваемых параметров. Это размер блока в обучающей выборке S_1, размер блока в исследуемом тексте S_2, количество блоков обучающей выборки K. Помимо этого, можно выделить неявные параметры, такие как способ сжатия данных и способ предобработки текста (например, можно убрать из текста всю пунктуацию или имена собственные).

Выбор параметров алгоритма

Для того, чтобы иметь возможность выбирать параметры алгоритма, необходимо определить способ измерения качества его работы. Тогда оптимальным значением параметра будет то, при котором качество будет наивысшим.

Измерить качество возможно, если в роли текста неизвестного автора на самом деле использовать текст, автор которого известен – в таком случае можно определить, ошибся алгоритм или нет. С этой целью среди текстов исследуемых авторов выделим тексты, авторов которых в рамках эксперимента будем считать неизвестными (тестовая выборка). Для этого в дополнение к обучающей выборке сформируем тестовую, т. е. для каждого *i*-го автора зафиксируем еще одно подмножество номеров блоков I_i^{test}, причем $\left|I_i^{test}\right| = K_2$, $I_i^{train} \cap I_i^{test} = \varnothing$. Для описания метода зафиксируем порядковый номер автора, которого обозначим A. Введем вектор $b_A = (0,0,...,0)$ размера N. Сожмем каждый блок с номером из I_A^{test} с обучающей выборкой каждого автора. Затем увеличим на единицу значение $b_{A,p}$, где p – номера автора, определенного по результату алгоритма как автор неизвестного блока. При корректной работе алгоритма в большинстве случаев должно выполняться равенство $\arg\max(b_A) = A$, т.е. в результате по большинству «наименьших сжатий» алгоритм определил в качестве автора неизвестного блока автора A. Заметим, что если проделать такую процедуру для каждого автора, то получим матрицу B, состоящую из строк b_i, $i \in \{1,...,N\}$.

Рассмотрим пример. Пусть имеются тексты произведений «Алые паруса», «Апельсины», «Белый огонь», «Ива» А. Грина (автор под номером 1); «Анафема», «Белый пудель», «Слон», «Яма» А. Куприна (автор под номером 2); «Игроки», «Ревизор», «Портрет», «Коляска» Н. Гоголя (автор под номером 3). Разделим каждое произведение ровно на два блока одинаковой длины. Первые половины произведениий включим в обучающую выборку каждого автора, в то время как вторые половины текста будут определены в тестовую. Схематичное отображение этого способа разбиения для произведений А. Грина приведено на рис. 1.

Руководствуясь алгоритмом, введем вектор $b_1 = (0,0,0)$, соответствующий автору A. Грину. Каждый тестовый блок указанного автора сожмем с обучающими текстами каждого автора. Пусть архиватор лучше всего сжал один блок с обучающим текстом А. Куприна, а оставшиеся три – с обучающим текстом А. Грина. Тогда $b_1 = (3,1,0)$. Аналогично, пусть бóльшая часть тестовых блоков А. Куприна лучше всего сжалась с обучающим текстом А. Куприна. В таком случае, например, $b_2 = (1,3,0)$. Наконец, пусть все тестовые блоки Н. Гоголя лучше всего сжались с обучающим текстом самого Н. Гоголя. Тогда $b_2 = (0,0,4)$. Объеденив полученные вектора в матрицу, получим:

$$B = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 4 \end{pmatrix}$$

Если для каждой строки b_i максимальное число находится на диагонали, т. е. выполняется $\arg\max(b_i) = i$, то алгоритм верно определяет автора (наблюдается «диагональное преобладание»). Мера этого преобладания определяет качество алгоритма: чем оно сильнее, тем лучше алгоритм справляется со своей задачей.

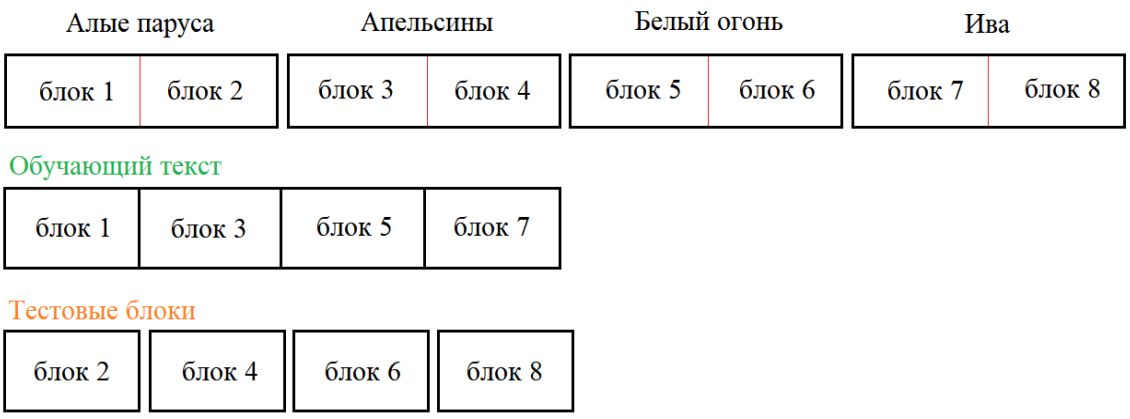


Рис. 1. Схематичное изображение разбиения произведений А. Грина на обучающую и тестовую выборки

Используя введенные ранее обозначения, формально запишем алгоритм в следующем виде:

$$B = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}.$$

Для $i = 1..N$

Для $p = 1..N$

$$C_p = \left| \sum_{u \in I_i^{train}} X_{pu} X_{ij} \right| - \left| \sum_{u \in I_i^{train}} X_{pu} \right|$$

$$B_{i,\arg\min(C)} += 1$$

(Здесь $\sum_{u \in I_i^{train}} X_{pu} X_{ij}$ – файл, полученный после объединения всех тренировочных блоков и тестируемого X_{ij}).

Если алгоритм определяет автора не лучше, чем просто случайный выбор, значения в матрице B будут распределены равномерно. Наоборот, если алгоритм способен определять настоящего автора, значения в матрице распределяться неоднородно (в случае идеальной работы алгоритма матрица B должна быть диагональной). Заметим, что в среднем величина статистики Хи-квадрат [18] для значений матрицы B будет тем больше, чем больше в матрице близких к нулю значений. Поэтому указанная статистика может служить мерой качества полученных результатов – чем она больше, тем точнее алгоритм определяет настоящего автора.

Выбор архиватора

Для того, чтобы выбрать оптимальный для рассматриваемого алгоритма архиватор, использованы тексты Б. Пастернака, Е. Замятина и А. Платонова. Как в обучающей, так и в тестовой выборке взяты все доступные в [19] произведения этих писателей. Было проведено несколько экспериментов для разных типов архиваторов: *BZIP2, DEFLATE, LZMA*.

Для каждого архиватора вычисления проводились 10 раз на разных обучающих и тестовых выборках. Объем обучающей выборки каждого автора – 512 Кб (~500 тыс. символов). Размер одного «приписываемого» блока – 4 Кб (~4 тыс. символов). Для каждого типа сжатия в табл. 1 приведена усредненная по всем экспериментам матрица B из указанного в предыдущем разделе алгоритма и вычисленная для нее статистика Хи-квадрат.

Из таблицы видно, что чем сильнее диагональное преобладание, тем больше величина Хи-квадрат. Ее наибольшее значение в среднем достигается для архиватора *LZMA*, поэтому в дальнейшем использовался еменно он.

Выбор способа предварительной обработки текста

Для дальнейших исследований была собрана значительная часть произведений писателей – современников И. Ильфа, Е. Петрова и М. Булгакова, написанных в период с 1915 по 1930 гг. При этом не использовались поэмы, стихотворения, сказки и детская литература.

Во многих работах, посвященных классификации и, в частности, атрибуции, авторы предлагают предварительно убирать из текстов числа, пунктуацию и так называемые стоп-слова (то есть слова, которые не несут смысловой нагрузки: частицы, союзы и т.д.) для того, чтобы снизить влияние на результат возможного текстового «шума» [17, 20]. Также многими авторами рекомендуется приведение слов к одному регистру.

Так как результаты классификации могут зависеть от предварительной обработки, нами также исследовался этот вопрос. В первую очередь мы удаляли из текстов лишние пробелы, переносы строки и нечитаемые символы – все то, что авторы точно не вносили в свои произведения самостоятельно. Кроме того, мы исследовали всевозможные комбинации предварительной обработки: с пунктуацией/без пунктуации, со стоп-словами/без стоп-слов, с заглавными буквами/без заглавных.

В качестве примера рассмотрим один из способов предварительной обработки. В табл. 2 приведены результаты для текстов, из которых была удалена пунктуация. Здесь наблюдается явное диагональное преобладание, а значит алгоритм успешно справляется с поставленной задачей при указанном способе предобработки текста. Аналогичные расчеты были проведены и для всех остальных перечисленных типов предобработки.

Значения статистики Хи-квадрат для каждого типа предобработки представлены в табл. 3. Видно, что оценка Хи-квадрат, полученная для текстов без заглавных букв, меньше, чем значение критерия для текстов с заглавными буквами. Других зависимостей не обнаружено, однако наибольшее значение критерия достигалось для текстов, из которых была удалена только пунктуация. В дальнейшем мы использовали этот вариант.

Таблица 1

Выбор архиватора: усредненная по всем экспериментам матрица *B*

	BZIP2			DEFLATE			LZMA		
	Пастернак	Замятин	Платонов	Пастернак	Замятин	Платонов	Пастернак	Замятин	Платонов
Пастернак	81	18	29	125	1	2	128	0	0
Замятин	19	85	24	16	107	5	2	124	2
Платонов	32	17	79	12	2	114	0	0	128
	Хи-квадрат:		203,3	Хи-квадрат:		571,2	Хи-квадрат:		743,4

Таблица 2

Матрица для текстов без пунктуации

	Пастернак	Набоков	Паустовский	Вересаев	Замятин	Бунин	Горький	Куприн	Платонов	Мандельштам	Олеша	Катаев	Грин	Житков	Эренбург	Тынянов	Гайдар	Федин
Пастернак	**13**	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
Набоков	0	**13**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Паустовский	0	1	**15**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Вересаев	0	0	0	**16**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Замятин	0	0	0	0	**13**	0	0	0	0	1	0	1	0	0	1	0	0	0
Бунин	0	0	1	0	0	**13**	0	1	0	0	0	0	1	0	0	0	0	0
Горький	0	0	0	0	0	0	**16**	0	0	0	0	0	0	0	0	0	0	0
Куприн	0	1	0	1	0	0	0	**10**	0	0	0	0	3	1	0	0	0	0
Платонов	0	0	0	0	0	0	0	0	**16**	0	0	0	0	0	0	0	0	0
Мандельштам	0	0	0	0	0	0	0	0	0	**15**	0	0	0	0	1	0	0	0
Олеша	0	0	0	0	0	0	0	0	0	1	**15**	0	0	0	0	0	0	0
Катаев	0	0	0	0	0	0	0	0	1	0	0	**15**	0	0	0	0	0	0
Грин	0	0	0	1	0	0	0	0	0	0	0	0	**15**	0	0	0	0	0
Житков	0	0	0	0	0	0	0	0	0	0	0	0	1	**15**	0	0	0	0
Эренбург	0	0	0	0	0	1	0	0	0	0	0	0	0	0	**15**	0	0	0
Тынянов	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**16**	0	0
Гайдар	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**16**	0
Федин	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	**14**

Таблица 3

Значения Хи-квадрат для разных типов предобработки текста

Тип предобработки	Хи-квадрат
Без пунктуации, без стоп-слов, без заглавных	3609,0
Без пунктуации, со стоп-словами, без заглавных	3789,0
Без пунктуации, без стоп-слов, с заглавными	4061,2
Без пунктуации, со стоп-словами, с заглавными	**4146,7**
С пунктуацией, без стоп-слов, без заглавных	3935.2
С пунктуацией, без стоп-слов, с заглавными	3903,7
С пунктуацией, со стоп-словами, без заглавных	3537,0
С пунктуацией, со стоп-словами, с заглавными	4133,2

Определение общего объема обучающей выборки и размера блока

В соответствии с алгоритмом текст неизвестного автора разбивается на блоки равного размера, а затем каждый из этих блоков дописывается к тексту, автор которого известен. Объем дописываемого блока влияет на качество работы метода, поэтому мы провели эксперименты, направленные на определение оптимального значения этой величины. Для одного и того же объема обучающей выборки, равного 64 Кб, варьировался размер дописываемого блока и для выбора этой величины проводились вычисления, аналогичные приведенным выше. В табл. 4 и 5 представлены результаты для блоков размерами соответственно 1 и 4 Кб.

Из табл. 4 и 5 и результатов вычисления статистики Хи-квадрат видно, что выбор размера блока не оказывает существенного влияния на полученные результаты. Это, в том числе, свидетельствует об устойчивости предлагаемого метода и о том, что он может быть успешно применен даже для сравнительно небольших текстов.

Базируясь на приведенных экспериментах, в дальнейших расчетах общий объем текста обучающей выборки мы установили равным 128 Кб, размер одного блока – 2 Кб, в качестве архиватора использовали *LZMA*, и в тексте убирали пунктуацию.

Таблица 4

Исследование влияния блока, имеющего размер 1 КБ

	Пастернак	Набоков	Паустовский	Вересаев	Замятин	Бунин	Горький	Куприн	Платонов	Мандельштам	Олеша	Катаев	Грин	Житков	Эренбург	Тынянов	Гайдар	Федин
Пастернак	**45**	0	1	0	0	1	0	3	0	2	1	0	3	0	4	1	0	3
Набоков	3	**33**	2	0	0	4	2	5	1	4	0	2	5	0	2	0	1	0
Паустовский	0	0	**41**	0	1	1	0	3	4	0	2	1	4	0	2	3	1	1
Вересаев	0	0	0	**58**	0	0	2	2	0	0	0	1	1	0	0	0	0	0
Замятин	2	0	1	0	**48**	0	0	2	3	1	1	0	1	0	3	0	0	2
Бунин	0	1	3	2	0	**46**	0	7	0	1	0	0	2	0	2	0	0	0
Горький	1	0	0	0	0	1	**54**	2	2	1	1	0	2	0	0	0	0	0
Куприн	1	2	1	3	1	0	1	**35**	1	2	0	3	7	0	5	0	1	1
Платонов	4	0	1	1	0	0	0	1	**54**	0	0	0	0	0	2	1	0	0
Мандельштам	0	0	0	1	0	2	0	2	1	**53**	0	1	0	0	3	0	0	1
Олеша	0	0	1	0	0	0	0	1	0	1	**58**	1	2	0	0	0	0	0
Катаев	0	0	0	0	0	0	0	1	1	0	0	**61**	0	0	0	0	1	0
Грин	0	2	3	4	1	0	0	5	0	0	0	0	**49**	0	0	0	0	0
Житков	0	0	2	1	1	0	0	1	0	0	0	0	5	**52**	0	0	2	0
Эренбург	1	1	2	0	0	1	1	1	1	3	0	1	2	0	**48**	1	1	0
Тынянов	0	0	0	0	0	0	0	3	0	1	1	0	0	0	0	**59**	0	0
Гайдар	0	0	0	2	0	0	0	0	0	0	0	1	3	0	0	0	**58**	0
Федин	5	0	1	1	1	0	0	1	1	7	0	0	0	0	3	0	0	**44**

Таблица 5

Исследование влияние блока, имеющего размер 4 КБ

	Пастернак	Набоков	Паустовский	Вересаев	Замятин	Бунин	Горький	Куприн	Платонов	Мандельштам	Олеша	Катаев	Грин	Житков	Эренбург	Тынянов	Гайдар	Федин
Пастернак	**15**	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Набоков	0	**12**	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0
Паустовский	0	0	**15**	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Вересаев	0	0	0	**16**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Замятин	0	0	0	0	**14**	0	0	0	0	1	0	1	0	0	0	0	0	0
Бунин	0	1	0	0	0	**12**	0	0	0	1	0	0	2	0	0	0	0	0
Горький	0	0	0	0	0	0	**16**	0	0	0	0	0	0	0	0	0	0	0
Куприн	0	0	0	0	0	0	1	**10**	0	0	0	1	4	0	0	0	0	0
Платонов	0	0	0	0	0	0	0	0	**16**	0	0	0	0	0	0	0	0	0
Мандельштам	0	0	0	0	0	0	0	0	0	**16**	0	0	0	0	0	0	0	0
Олеша	0	0	0	0	0	0	0	0	0	0	**16**	0	0	0	0	0	0	0
Катаев	0	0	0	0	0	0	0	0	0	0	0	**16**	0	0	0	0	0	0
Грин	0	0	0	0	0	0	0	0	0	0	0	0	**16**	0	0	0	0	0
Житков	0	0	0	0	0	0	0	0	0	0	0	0	0	**16**	0	0	0	0
Эренбург	0	0	0	0	0	0	0	1	0	2	0	0	1	0	**12**	0	0	0
Тынянов	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**16**	0	0
Гайдар	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	**15**	0
Федин	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	**15**

АНАЛИЗ ПРОИЗВЕДЕНИЙ «12 СТУЛЬЕВ» И «ЗОЛОТОЙ ТЕЛЕНОК»

Как отмечалось ранее, в последнее время получила распространение гипотеза о том, что произведения «12 стульев» и «Золотой теленок» написал М. Булгаков, а не И. Ильф и Е. Петров. В этом разделе мы проверим эту гипотезу при помощи описанного выше метода.

Сначала протестируем наш метод на произведении, чье авторство является бесспорным. В табл. 6 представлены результаты применения исследуемого метода. В качестве обучающей выборки использовалось подмножество всех текстов И. Ильфа и Е. Петрова общим объемом 128 Кб и подмножество всех произведений М. Булгакова, из которых было исключено «Собачье сердце». В качестве тестовой выборки использовались блоки произведения «Собачье сердце» размером 2 Кб. Общий объем тестовой выборки – 128 Кб (т. е. в тестовой выборке было 64 блока).

На основании табл. 6 можно сделать однозначный вывод об авторстве произведения «Собачье сердце», что в данном случае свидетельствует о качестве метода.

Проведем исследование отделимости стилей рассматриваемых авторов. Не меняя обучающую выборку, в качестве тестовой возьмем по 128 Кб произвольных текстов каждого из авторов. В табл. 7 представлены результаты этого эксперимента.

Из табл. 6 и 7 видно, что авторы обладают уникальными, отличимыми стилями. А значит авторство и других написанных ими произведений также должно однозначно определяться.

Наконец, выясним авторство произведений «12 стульев» и «Золотой теленок». В табл. 8 представлены результаты вычислений.

Судя по этим результатам авторы «12 стульев» и «Золотой теленок» – И. Ильф и Е. Петров. Однако, в отличие от эксперимента с произведением «Собачье сердце», часть блоков все же была отнесена к М. Булгакову. По-видимому, исследуемые произведения действительно в какой-то мере похожи на произведения М. Булгакова.

Наконец, мы провели исследование близости стилей этих двух произведений к стилям наиболее известных современников И. Ильфа и Е. Петрова. Результаты приведены в табл. 9, из которой видно, что и «Золотой теленок», и «12 стульев», вероятнее всего, были написаны И. Ильфом и Е. Петровым. Более того, сходство «Золотого теленка» с произведениями К. Паустовского выше, чем с произведениями Булгакова, (что подтверждается результатами из табл. 10 и 11), хотя авторы обладают вполне уникальными, отличимыми стилями, что видно из табл. 10.

По нашему мнению, приведенные данные однозначно свидетельствуют об авторстве И. Ильфа и Е. Петрова, однако они показывают и заметное сходство со стилем М. Булгакова и К. Паустовского.

Возможно, это связано с тем, что во время журналистской работы в Одессе К. Паустовский очень сблизился с И. Ильфом и И. Бабелем [21]. В ранних рассказах Паустовского, окрашенных романтизмом, угадываются интонации Бабеля [22]. Такое же влияние мог оказать на его творчество Илья Ильф. Тем более, что в 1928 г. К. Паустовский издал роман «Блистающие облака», в детективно-авантюрную сюжетную линию которого вплетались автобиографические эпизоды, описывающие поездки К. Паустовского по Кавказу и побережью Черного моря в 1925–1927 гг. В том же 1928 г. был издан роман «12 стульев». Романы создавались одновременно, а именно зимой 1927 г. и были написаны в рекордно короткие сроки, будто авторы соревновались друг с другом. Важно напомнить, что И. Ильф и К. Паустовский были близкими друзьями со времен совместной журналистской деятельности в Одессе. Возможно, этим объясняется то, что исследуемые произведения чем-то похожи на произведения К. Паустовского.

Таблица 6

Атрибуция произведения "Собачье сердце"

	М. Булгаков	Е. Петров и И. Ильф
Собачье сердце	64	0

Таблица 7

Анализ текстов исследуемых авторов

	М. Булгаков	Е. Петров и И. Ильф
М. Булгаков	**63**	1
Е. Петров и И. Ильф	0	**64**

Таблица 8

Атрибуция произведений «12 стульев» и «Золотой теленок»

	М. Булгаков	Е. Петров и И. Ильф
12 стульев	7	57
Золотой теленок	5	59

Таблица 9

Оценка близости стиля авторов произведений «12 стульев» и «Золотой теленок» к стилям других авторам

	Булгаков	Ильф и Петров	Замятин	Пастернак	Паустовский	Чехов	Платонов	Набоков	Вересаев	Бунин	Горький	Куприн	Мандельштам	Олеша	Катаев	Грин	Житков	Эренбург	Тынянов	Гайдар	Федин
Золотой теленок	4	**13**	0	2	7	0	1	0	0	0	0	2	0	0	1	0	0	1	**1**	0	0
12 стульев	5	**9**	1	2	3	0	1	0	0	0	0	1	2	0	2	0	0	2	3	**0**	1

Таблица 10

Проверка близости стиля произведений различных авторов к стилю романов «Золотой теленок» и «12 стульев»

	Булгаков	Ильф и Петров	Паустовский
Булгаков	**28**	3	1
Ильф и Петров	2	**29**	1
Паустовский	0	3	**29**

Таблица 11

Исследование близости стиля романов «12 стульев» и «Золотой теленок» к стилям М. Булгакова, К. Паустовского, И. Ильфа и Е. Петрова

	Булгаков	Ильф и Петров	Паустовский
Золотой теленок	2	21	9
12 стульев	8	16	8

ЗАКЛЮЧЕНИЕ

Нами были рассмотрены формальные методы классификации текстов на естественных языках и, в частности, проблемы атрибуции. Формальные методы не заменяют и, конечно, не отменяют традиционные подходы, основанные на работе филологов, литературоведов и историков. Скорее, формальные и «традиционные» методы должны применяться вместе и дополнять друг друга.

При классификации и атрибуции исследователи работают с очень большими объемами информации, следствием чего является необходимость учитывать частоты использования слов и их сочетаний разными авторами. Количество сочетаний слов растет в геометрической прогрессии и уже при фразе из трех-четырех слов становится астрономическим. Более того, кроме оценок встречаемости слов и фраз следует учитывать такие трудно формализуемые характеристики, как стиль автора и произведения, создаваемое им настроение и многое другое. В целом задачи классификации и атрибуции литературных произведений можно считать разновидностью проблемы анализа «больших данных» (*Big Data challenge*). Их решение требует разработки специальных методов, обычно относимых к области искусственного интеллекта, и заведомо превосходящих возможности человека оперировать с очень большими объемами данных. В связи с этим важно отметить, что методы сжатия данных, играющие в нашем исследовании ключевую роль, основаны на опыте их разработчиков при анализе больших объемов данных, а также на применении методов теории информации, формальных грамматик и ряда других подходов теории искусственного интеллекта. По-видимому, именно по этой причине методы сжатия данных находят все более широкое применение в задачах обработки текстов.

В нашей работе подход к классификации текстов на основе использования методов сжатия данных адаптирован и «настроен» для рассматриваемой задачи определения авторства романов «12 стульев» и «Золотой теленок». При этом сначала разработанный метод протестирован на текстах писателей – современников И. Ильфа, Е. Петрова и М. Булгакова, и было показано, что он способен успешно справляться с решением задач атрибуции.

Полученные в работе результаты свидетельствуют о том, что авторами романов «12 стульев» и «Золотой теленок» являются И. Ильф и Е. Петров. Однако, исходя из результатов исследования, можно сказать, что стиль этих романов несколько напоминает произведения М. Булгакова и К. Паустовского. Интересно, что влияние

М. Булгакова отмечалось некоторыми исследователями творчества И. Ильфа и Е. Петрова, и объясняется, возможно, тем, что он мог принимать участие в обсуждении романов в период их написания. Некоторое сходства стилей этих двух романов и произведений К. Паустовского ранее не отмечалось и, возможно, вызовет интерес у литературоведов.

СПИСОК ЛИТЕРАТУРЫ

1. Lowe D., Matthews R. Shakespeare vs. Fletcher: A Stylometric Analysis by Radial Basis Functions // Computers and the Humanities. – 1995. – Vol. 29, № 26. – P. 449-461.
2. Марусенко М.А. Атрибуция анонимных и псевдонимных литературных произведений методами теории распознавания образов. – Л.: Изд-во Ленингр. ун-та, 1990. 164 с.
3. Марусенко М.А., Бессонов Б.Л., Богданова Л.М., Аникин М.А., Мясоедова Н.Е. В поисках потерянного автора. Этюды атрибуции. – СПб: Филол. фак. С.-Петерб. гос. ун-та, 2001. – 209 с.
4. Милов Л.В., Бородкин Л.И., Иванова Т.И. и др. От Нестора до Фонвизина. Новые методы определения авторства. – М: Прогресс, 1994.
5. Мовсисян А. Марк Леви: Роман с Ереваном // Газета.РАУ. – 2015. – №184 (10.04.2015). – URL: http://gazeta.rau.am/index.php?bid=1203 (дата обращения: 17.03.2019).
6. Амлински И. 12 стульев от Михаила Булгакова. – Берлин: Kirschner Verlag, 2013. – 328 с.
7. Козаровецкий В.А. Московские Баранки и Одесские бублики. Кто написал «12 стульев» // Литературная Россия. – 2013. – №41 (11.10.2013). – URL: http://old.litrossia.ru/2013/41/08347.html (дата обращения: 10.10.2018)
8. Хмельницкий Д.С. В защиту Ильфа и Петрова // Семь искусств. – 2014. – №52 (13.05.2014). – URL:http://7iskusstv.com/2014/Nomer5/Chmelnicky1.php (дата обращения: 18.03.2019)
9. Фрейдгейм Л.И. Ильф и Петров или Булгаков… Круглый стол (виртуальный вариант) // Семь искусств. – 2014. – №47 (13.11.2013). – URL: http://7iskusstv.com/2013/Nomer11/Frejdgejm1.php (дата обращения: 18.03.2019)
10. Юрганов А.А. Авторский инвариант как метод атрибуции текста // Проблемы Науки. – 2017. – №29(111). – URL: https://cyberleninka.ru/article/n/avtorskiy- invariant-kak-metod-atributsii-teksta (дата обращения: 15.02.2019).
11. Хмелёв Д.В. Распознавание автора текста с использованием цепей А.А. Маркова // Вестник МГУ. – 2000. – № 2. – С. 115-126
12. Кукушина О.В., Поликарпов А.А., Хмелёв Д.В. Определение авторства текста с использованием буквенной и грамматической информации // Проблемы передачи информации». – 2001. – Т. 37, № 2. – С. 96-109
13. Gorshkov S., Nered M., Ilyushin E., Namiot D. Using machine learning methods to establish program authorship // International Journal of Open Information Technologies. – 2019. – №1. – URL: https://cyberleninka.ru/article/n/using-machine-learning-methods-to-establish-program-authorship (дата обращения: 14.02.2019).
14. Sidorov G.O. Automatic authorship attribution using syllables as classification features // Rhema. Рема. – 2018. – №1. – URL: https://cyberleninka.ru/article/n/automatic-authorship-attribution-using-syllables-as-classification-features (дата обращения: 14.02.2019).
15. Хмелёв Д.В. Классификация и разметка текстов с использованием методов сжатия данных // Все о сжатии данных, изображений и видео. – 2003. – URL: http://compression.ru/download/articles/ classif/intro.html
16. Lalli M., Tria F., Loreto V. Data-Compression Approach to Authorship Attribution // Workshop Proceedings. September, 2017, Padova: Padova University Press. – P. 61.
17. Ryabko B., Astola J., Malyutov M. Compression-based methods of statistical analysis and prediction of time series. – New York: Springer International Publishing, 2016.
18. Кендалл М., Стюарт А. Статистические выводы и связи: пер. с англ. – М: Наука. Гл. ред. физ.-мат. лит., 1973.
19. Библиотека Максима Мошкова. – URL: http://lib.ru
20. Селиванова И.В., Рябко Б.Я., Гуськов А.Е. Классификация посредством компрессии: применение методов теории информации для определения тематики научных текстов // Научно-техническая информация. Сер. 2. – 2017. – №6. – С. 8-15; Selivanova I.V., Ryabko B.Ya., Guskov A.E. Classification by compression: Application of information-theory methods for the identification of themes of scientific texts // Automatic Documentation and Mathematical Linguistics. – 2017. – Vol. 51, № 3. – P. 120–126.
21. Паустовский К. Время больших ожиданий: в 2 т. – М.: Изд-во Деком, 2002.
22. Трефилова Г. О стиле Паустовского // Новый мир. – 1967. – № 4. – С. 214-220.

Материал поступил в редакцию 18.04.19.

Сведения об авторах:

ПОЛОХИН Степан Николаевич – магистрант, факультет информационных технологий, Новосибирский государственнй университет
e-mail: ilikebigcat@yandex.ru

РЯБКО Борис Яковлевич – доктор технических, профессор, Институт вычислительных технологий СО РАН, главный научный сотрудники заведующий лабораторией информационных систем и защиты информации; Новосибирский государственный университет, профессор.
e-mail: boris@ryabko.net

САВИНА Надежда Николаевна – доктор педагогических наук, доцент; Новосибирский государственный университет, ведущий научный сотрудник лаборатории теоретико-информационных методов;
e-mail: nn_savina@mail.ru

УВАЖАЕМЫЕ КОЛЛЕГИ!

ВИНИТИ РАН предлагает Вашему вниманию Реферативный Журнал в электронной форме

РЖ в электронной форме (ЭлРЖ) выпускается по всем разделам естественных, технических и точных наук.

Каждый номер ЭлРЖ является полным аналогом печатного номера РЖ по составу описаний документов, их оформлению и расположению. Он сопровождается оглавлением, указателями.

ЭлРЖ представляет собой информационную систему, снабженную поисковым аппаратом и позволяющую пользователю на персональном компьютере:

- читать номер РЖ, последовательно листая рефераты;
- просматривать рефераты отдельных разделов по оглавлению;
- обращаться к рефератам по указателям авторов, источников, ключевых слов;
- проводить поиск документов по словам и словосочетаниям;
- выводить текст описаний документов во внешний файл.

ЭлРЖ в версии Windows Вы можете получить за текущий год с любого номера, а также за предыдущие годы.

Подробную информацию Вы можете получить:

Адрес: 125190, Россия, Москва, ул. Усиевича, 20, ВИНИТИ РАН

Телефон 499-155-42-85 499-151-78-61

E-mail: Contact@viniti.ru, Feo@viniti.ru

ВНИМАНИЮ ЧИТАТЕЛЕЙ!

ИЗДАНИЕ УДК

УНИВЕРСАЛЬНАЯ ДЕСЯТИЧНАЯ КЛАССИФИКАЦИЯ

АЛФАВИТНО-ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

в 2-х томах

Алфавитно-предметный указатель (АПУ) к 4-му полному изданию УДК на русском языке:

Том I содержит АПУ от буквы А до Н;

Том II содержит АПУ от буквы М до Я и указатель латинских наименований к классам УДК 56 Палеонтология, 57 Биологические науки, 58 Ботаника, 49 Зоология, 61 Медицинские науки.

АПУ содержит около 100 000 понятий, представленных в полных таблицах УДК.

При его составлении были учтены изменения, опубликованные в Выпусках № 1 – 6 «Изменения и дополнения к УДК»

Для подписки необходимо направить заявку для оформления счета по адресу:

125190, Россия, Москва, ул. Усиевича, 20, ВИНИТИ РАН

Телефоны: 499 155-42-85, 499 151-78-61

E-mail: feo@viniti.ru

http://www.udcc.ru