ФЕДЕРАЛЬНОЕ АГЕНТСТВО НАУЧНЫХ ОРГАНИЗАЦИЙ

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

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 5 Москва 2015

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ

УДК 510.64 : 004.89

С.М. Гусакова

Корректность ДСМ-рассуждений для однородных стратегий*

Рассматривается определение корректности ДСМ-рассуждений для различных методов и стратегий. Показано, что экстенсиональная решетка методов с отношением порядка по вложению множеств гипотез индуцирует изоморфную решетку однородных стратегий, отношение порядка в которой определяется следованием тотальной корректности для одной стратегии из другой.

Ключевые слова: *корректность ДСМ-рассуждений, отношение толерантности, транзитивность, тотальная корректность, экстенсиональная решетка*

ДСМ-метод автоматического порождения гипотез является не только инструментом для выдвижения гипотез о причинах явлений и свойствах объектов, но и дает возможность определить корректность ДСМ-рассуждений и на основании этого делать выводы о том, является ли полученная эмпирическая зависимость закономерностью или тенденцией [1].

Корректность ДСМ-рассуждений связана с непротиворечивостью множеств гипотез, получаемых в процессе работы ДСМ-метода при последовательном расширении баз фактов (БФ).

Поскольку ДСМ-метод формализует различные индуктивные методы Д.С. Милля [2], то на одной и той же последовательности расширяющихся баз фак-

* Работа выполнена при поддержке РФФИ (проект № 15-07 -02 402) и Программы фундаментальных исследований РАН на 2015 г. (№ П8).

тов корректность ДСМ-рассуждения зависит от применяемых методов.

Полный цикл ДСМ-рассуждения на заданной базе фактов называется этапом. Непротиворечивость определяется с помощью некоторых функционалов, которые в свою очередь определяют на множестве этапов отношение толерантности. В работе исследуется это отношение и его связь с корректностью ДСМ-рассуждений.

Определим основные понятия. Полный цикл применения тактов (индукция → аналогия)$_1$ →(индукция → аналогия)$_2$→ … (индукция → аналогия)$_n$ такой, что множество порожденных гипотез на такте n совпадает с множеством гипотез, порожденных на такте $n + 1$, где n – номер первого такого совпадения, составляет этап I ДСМ-рассуждения [3]. Когда этап I завершен, после проведения некоторых процедур, связанных с абдуктивным рассуждением, образующих этап II (см. [1]), происходит расширение базы фактов и переход к этапу I со следующим номером. В результате этого процесса возникает множество этапов $\{Ik\}$.

ДСМ-рассуждения можно проводить с помощью различных методов. Поскольку при применении правил правдоподобного вывода для получения гипотез о причинах используются как положительный, так и отрицательный методы , то мы получаем различные стратегии $Str_{x,y}$, где x – положительный, а y – отрицательный методы. В нашей статье рассматриваются только однородные стратегии, т. е. такие, у которых положительный и отрицательный методы однотипны.

Через Δ_p^{σ} , где $\sigma \in \{+, -, 0\}$ обозначается множество гипотез о причинах (гипотезы I рода) вида $J_{\langle\sigma,l\rangle}(C \Rightarrow_2 Q)$, полученных на этапе I с номером p. При этом Δ_p^{+} – множество положительных, Δ_p^{-} – множество отрицательных, а Δ_p^{0} – множество противоречивых гипотез. $C \Rightarrow_2 Q$ называется телом гипотезы.

Рассмотрим выражение: $\Delta_p^{\sigma_1} \cap (\Delta_q^{\sigma_2} \cup \Delta_q^{\sigma_3})$, где $\sigma_1, \sigma_2, \sigma_3 \in \{+, -, 0\}$ и σ_1, σ_2 и σ_3 – различны, p и q – номера этапов. По сути это три выражения: $\Delta_p^{+} \cap (\Delta_q^{-} \cup \Delta_q^{0})$, $\Delta_p^{-} \cap (\Delta_q^{+} \cup \Delta_q^{0})$ и $\Delta_p^{0} \cap (\Delta_q^{+} \cup \Delta_q^{-})$. При этом операция $\cup$ определяется стандартным образом, а операция $\cap$ определяется следующим образом: если $J_{\langle 1,l\rangle}(C \Rightarrow_2 Q) \in \Delta_p^{+}$, а $J_{\langle \nu,k\rangle}(C \Rightarrow_2 Q) \in \Delta_q^{-} \cup \Delta_q^{0}$, где $\nu \in \{-1, 0\}$, то $C \Rightarrow_2 Q \in \Delta_p^{+} \cap (\Delta_q^{-} \cup \Delta_q^{0})$. Аналогично определяется $\cap$ для $\Delta_p^{-} \cap (\Delta_q^{+} \cup \Delta_q^{0})$ и $\Delta_p^{0} \cap (\Delta_q^{+} \cup \Delta_q^{-})$.

Множество номеров этапов I на последовательности расширяющихся баз фактов обозначим через N. Определим на множестве N отношение строгого порядка $<$:

$$p < q \rightleftharpoons БФ_p \subset БФ_q.$$

Определим также на этом множестве два бинарных отношения $\tilde{R}(p,q)$ и $\tilde{K}(p,q)$ следующим образом:

$$\tilde{R}(p,q) \rightleftharpoons \Delta_p^{+} \cap (\Delta_q^{-} \cup \Delta_q^{0}) = \varnothing \;\&\; \Delta_p^{-} \cap (\Delta_q^{+} \cup \Delta_q^{0}) = \varnothing \;\&\; \Delta_p^{0} \cap (\Delta_q^{+} \cup \Delta_q^{-}) = \varnothing.$$

$$\tilde{K}(p,q) \rightleftharpoons \Omega^{+}{}_p \cap (\Omega^{-}{}_q \cup \Omega^{0}{}_q) = \varnothing \;\&\; \Omega^{-}{}_p \cap (\Omega^{+}{}_q \cup \Omega^{0}{}_q) \;\&\; \Omega^{0}{}_p \cap (\Omega^{+}{}_q \cup \Omega^{-}{}_q).$$

Здесь Ω^{+} , Ω^{-} , Ω^{0} — гипотезы доопределения (гипотезы второго рода) вида $J_{\langle \nu,n\rangle}(C \Rightarrow_1 Q)$, $\nu \in \{+, -, 0\}$.

В [3] показано, что отношения $\tilde{R}$ и $\tilde{K}$ – рефлексивны и симметричны, следовательно, являются отношением толерантности [4]. Выполнение этих отношений означает, что множества положительных, отрицательных и противоречивых гипотез, полученных на этапе с номером p, не вступают в противоречие с объединенным множеством гипотез другого знака, полученных на этапе с номером q.

Итак, на множестве N определены три отношения – отношение строгого порядка $<$ и отношения толерантности $\tilde{R}$ и $\tilde{K}$.

В [3] было введено понятие тотальной корректности ДСМ-рассуждения. А в [5] – понятие тотальной корректности ДСМ-рассуждения I рода.

Определение 1. ДСМ-рассуждение тотально корректно, если для любых p и q таких, что $1 \le p, q \le s$ выполнено $\tilde{R}(p,q) \;\&\; \tilde{K}(p,q)$.

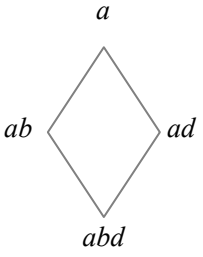
Определение 2. ДСМ-рассуждение обладает тотальной корректностью I рода, если $\forall\, p$ и q таких, что $1 \le p, q \le s$ выполнено $\tilde{R}(p,q)$.

Определение 3. ДСМ-рассуждение обладает тотальной корректностью II рода, если $\forall\, p$ и q выполнено $\tilde{K}(p,q)$, $1 \le p < q \le s$.

Если для любых p и q таких, что $1 \le p, q \le s$ имеет место $\neg\, \tilde{R}(p,q)$, то можно говорить о тотальной противоречивости I рода. Она свидетельствует о неудачно подобранных базах фактов или о том, что операция сходства не отражает суть решаемой задачи [6].

Тотальная противоречивость II рода имеет место при $\neg\, \tilde{K}(p,q)$ $\forall\, p$ и q таких, что $1 \le p, q \le s$. Этот вид противоречивости указывает на неадекватную модель предметной области.

Рассмотрим экстенсиональную решетку (см. [7]) однородных стратегий вида:



Рисунок

Здесь *a, ab, ad,abd* – имена предикатов простого сходства, простого сходства с запретом на контрпримеры, различия и различия с запретом на контрпримеры соответственно.

Отношение порядка в решетке определяется по вложению множеств гипотез *I* рода.

Исследуем, как устроены пространства толерантности (N, $\widetilde{R}$) на множестве номеров этапов N для однородных стратегий с предикатами из этой решетки.

В [5] было показано, что для простого метода сходства имеет место:

Утверждение 1. ДСМ-рассуждение обладает тотальной корректностью *I*-го рода, тогда и только тогда, когда $\forall p$ выполнено $\widetilde{R}$ (p, s), $1 \le p \le s$.

Следовательно, чтобы построить пространство толерантности (N, $\widetilde{R}$), достаточно проверить непротиворечивость каждого этапа только с последним, а не с каждым.

В [5] было исследовано также (N, $\widetilde{R}$) для простого метода сходства с запретом на контрпримеры. Запрет на контрпримеры выражает тот факт, что причина V и ее следствие W не содержатся ни в одном примере противоположного знака. Предикаты простого положительного и отрицательного сходства с запретом на контрпримеры обозначаются через $M^{+}_{ab,n}(V,W)$ и $M^{-}_{ab,n}(V,W)$ – соответственно.

Если используется метод простого положительного и отрицательного сходства с запретом на контрпримеры, то при расширении базы фактов могут появиться контрпримеры для гипотез, полученных на предыдущих этапах, тогда такие гипотезы исчезают при переходе на следующий этап. Следовательно противоречивые гипотезы при такой стратегии образоваться не могут, так как примеры противоположного знака, вступающие в конфликт, являются друг для друга контрпримерами. Поэтому имеет место $\forall p\ \forall q$ $\widetilde{R}$ (p, q), $(1 \le p, q \le s)$.

Утверждение 2.

Если ДСМ-рассуждение проводится с помощью стратегии с простым предикатом сходства с запретом на контрпримеры как для положительного, так и для отрицательного предикатов, то для такого рассуждения пространство толерантности (N, $\widetilde{R}$) всегда является полным. т. е. имеет место тотальная корректность *I* рода.

Рассмотрим метод различия, который определяется следующим образом:

$$M^{\pm}_{ad,n}(V,W) \rightleftharpoons M^{\pm}_{a,n}(V,W)\ \&\ (d^{\pm}),$$

где

$$(d^{+}) \rightleftharpoons \forall X \forall Y \forall Z \forall U ((J_{(1,n)}(X \Rightarrow_1 Y)\ \&\ (W \subseteq Y)\ \&\ (V \subset X)\ \&\ ((X \setminus V) \subset Z)\ \&\ ((X \setminus V) \neq \varnothing)\ \&\ \neg (V \subset Z)) \rightarrow (\neg J_{(1,n)}(Z \Rightarrow_1 U) \rightleftharpoons \neg (W \subseteq U))),$$

а d^{-} определяется аналогично.

Очевидно, что множество гипотез *I* рода, удовлетворяющих (±) предикатам простого сходства, содержит множество гипотез *I* рода, удовлетворяющих (±)-предикатам различия [7]. Поэтому, если $\widetilde{R}$ (p,q) имеет место для $M^{\pm}_{a,n}$, то и для $M^{\pm}_{a,n}\ \&\ (d^{\pm})$. Следовательно ($N$, $\widetilde{R}$) для стратегии с простым методом сходства является подпространством (N, $\widetilde{R}$) для стратегии с методом различия, а это значит, что из тотальной корректности *I* рода для простого метода сходства вытекает и тотальная корректность *I* рода для метода различия.

При этом наличие противоречий между этапами в методе простого сходства может не помешать тотальной корректности метода различия. В самом деле, гипотеза $J_{\langle 1,n\rangle}(C \Rightarrow_2 Q)$, полученная простым методом сходства на этапе p и вступившая в конфликт с гипотезой $J_{\langle -1,n\rangle}(C \Rightarrow_2 Q)$, полученной на этапе q, в методе различия может отсутствовать в силу того, что она не удовлетворяет условию d^{+}.

Метод различия может быть усилен запретом на контрпримеры. Этот метод имеет вид: $M^{+}_{ab,n}\ \&\ d^{+}$ и $M^{-}_{ab,n}\ \&\ d^{-}$ и обозначается через $M^{+}_{abd,n}$ и $M^{-}_{abd,n}$.

Гипотезы *I* рода, полученные методом различия с запретом на контрпримеры, вкладываются в множество гипотез *I* рода, полученных методом простого сходства с запретом на контрпримеры. Поэтому из *Утверждения 2* следует:

Утверждение 3.

Если ДСМ-рассуждение проводится с помощью стратегии с предикатом различия с запретом на контрпримеры как для положительного, так и для отрицательного предикатов, то для такого рассуждения всегда имеет место тотальная корректность *I* рода.

Все полученные утверждения позволяют сделать вывод о том, что экстенсиональная решетка, приведенная на Рисунке, индуцирует изоморфную решетку для однородных стратегий с предикатами простого сходства, сходства с запертом на контрпримеры, различия и различия с запретом на контрпримеры. Отношение нестрого частичного порядка определяется в этой решетке следующим образом: $Str_i \ge Str_j$ тогда и только тогда, когда из тотальной корректности *I* рода для Str_i следует тотальная корректность *I* рода для Str_j.

Помимо методов, представленных в решетке на Рисунке, в ДСМ-рассуждениях используется еще и обратный метод сходства.

Предикаты обратного метода имеют вид:

$$\breve{M}^{+}_{a,n}(V, W) \rightleftharpoons \exists k\ \tilde{\breve{M}}^{+}_{a,n}(V,W,k), \text{ где}$$

$$\tilde{\breve{M}}^{+}_{a,n}(V,W,k) \rightleftharpoons$$

$$\exists X_1 ... \exists X_k . \exists Y_1 ... \exists Y_k (\&_{h=1}^{k}(J_{(+1,n)}(X_h \Rightarrow_1 Y_h)))$$

$$\&\ (\bigcap_{h=1}^{k} X_h = V)\ \&\ (V \neq \varnothing)\ \&\ (\bigcap_{h=1}^{k} Y_h = W)\ \&$$

$$(W \neq \varnothing)\ \&\ \forall i \forall j\ (((i \neq j)\ \&\ (1 \le i, j \le k) \rightarrow (X_i \neq X_j))\ \&$$

$$\forall X \forall Y ((J_{(+1,n)}(X \Rightarrow_1 Y))\ \&\ (W \subseteq Y))) \rightarrow$$

$$((V \subset X)\ \&\ \vee_{h=1}^{k}(Y = Y_h)))\ \&\ (k \ge 2)).$$

$$\breve{M}^{-}_{a,n}(V, W) \rightleftharpoons \exists k\ \tilde{\breve{M}}^{-}_{a,n}(V,W,k), \text{ где}$$

$$\tilde{\breve{M}}^{-}_{a,n}(V,W,k) \rightleftharpoons \exists X_1 ... \exists X_k . \exists Y_1 ... \exists Y_k$$

$$(\&_{h=1}^{k}(J_{(-1,n)}(X_h \Rightarrow_1 Y_h)))\ \&$$

$$(\bigcap_{h=1}^{k} X_h = V)\ \&\ (V \neq \varnothing)\ \&\ (\bigcap_{h=1}^{k} Y_h = W)\ \&$$
$$(W \neq \varnothing)\ \&\ \forall i \forall j\ (((i \neq j)\ \&\ (1 \leq i, j \leq k) \rightarrow (X_i \neq X_j)\ \&$$
$$\forall X \forall Y\ ((J_{(-1,n)}(X \Rightarrow_1 Y))\ \&\ (W \subseteq Y)\ ))) \rightarrow$$
$$((V \subset X)\ \&\ \vee_{h=1}^{k}(Y = Y_h))))\ \&\ (k \geq 2)).$$

Предикат $W_3 \Leftarrow V$ читается как «W есть следствие V».

Основное отличие обратного метода сходства от прямого состоит в том, что исчерпываемость примеров определяется по правым, а не по левым частям. Эта особенность обратного метода вносит существенные изменения в вопрос, связанный с непротиворечивостью ДСМ-рассуждений.

Предположим, что Δ_p^+ содержит гипотезу $W_3 <= V$. Поскольку при расширении базы фактов на этапе $p+1$ может появиться положительный пример $J_{<1,0>}(X_{k+1} \Rightarrow_1 Y_{k+1})$ такой, что $\bigcap_{h=1}^{k+1} Y_h = W$, а $\bigcap_{h=1}^{k+1} X_h \neq V$, то в Δ_{p+1}^+ этой гипотезы уже не будет. Если гипотеза с таким же телом появится в Δ_q^-, где $q>p$, будет иметь место $\neg\ \widetilde{R}\ (p, q)$. Таким же образом гипотеза $W_3 \Leftarrow V$ может исчезнуть и из Δ_t^- на этапе I с номером t, $t > q$. Тогда p и t будут толерантны. Таким образом, в силу особенности обратного метода соотношения между множествами гипотез Δ_p^{σ} и $\Delta_q^{\sigma_1}$, где $\sigma, \sigma_1 \in \{+, -, 0\}$, $\sigma \neq \sigma_1$ этапов Ip и Iq может быть произвольным, т. е. возможны варианты:

$$\Delta_p^{\sigma} \subseteq \Delta_q^{\sigma_1} \quad \text{или} \quad \Delta_q^{\sigma_1} \supseteq \Delta_p^{\sigma},$$
$$\Delta_p^{\sigma} \cap \Delta_q^{\sigma_1} = \varnothing,$$
$$\Delta_p^{\sigma} \cap \Delta_q^{\sigma_1} \neq \varnothing,$$

следовательно $\forall\ p, q\ (1 \leq p, q \leq s)$ может быть как $\widetilde{R}\ (p,q)$, так и $\neg\ \widetilde{R}\ (p, q)$ и на множестве N порождается произвольное пространство толерантности.

Для прямого и обратного методов сходства существует усиление в виде условия единственности причины $\forall Z\ (M_{a,n}^{-}(Z,W) \rightarrow (Z=V))$ для прямого метода и условия единственности следствия: $\forall U(J_{(1,n)}(U_3 \Leftarrow V) \rightarrow (U=W)$ для обратного метода.

Если V – единственная положительная причина следствия W на этапе с номером p, то при появлении на этапе с номером $p+1$ новых положительных причин этого следствия, множество положительных причин этапа $p+1$ окажется пустым в силу условия единственности причины. И если при этом в расширенной базе данных появилась единственная отрицательная причина V для следствия W, то $\Delta_p^{\sigma} \cap \Delta_q^{\sigma_1} \neq \varnothing$.

Аналогично для причин других знаков. Подобное рассуждение можно провести и для обратного метода, усиленного условием единственности следствия.

Таким образом для простого метода сходства с условием единственности причины и для обратного метода, а также обратного метода, усиленного условием единственности следствия проверка тотальной корректности I рода требует проверки выполнения $\widetilde{R}\ (p, q)\ \forall\ 1 \leq$ p, q $\leq$ s.

Поскольку тотальная корректность I рода означает, что пространство толерантности (N, $\widetilde{R}$) является полным пространством, т. е. $\widetilde{R}$ – транзитивно, имеет смысл исследовать, при каких условиях отношение $\widetilde{R}$ будет транзитивным.

Утверждение 4. Для того чтобы отношение толерантности $\widetilde{R}$ на множестве N было транзитивным, достаточно, чтобы $\forall\ p,q,t$ таких, что $\widetilde{R}\ (p,q)$ и $R\ (q,t)$, имело место:

$$(\Delta_p^{\sigma} \subseteq \Delta_q^{\sigma} \subseteq \Delta_t^{\sigma}) \vee (\Delta_p^{\sigma} \supseteq \Delta_q^{\sigma} \supseteq \Delta_t^{\sigma}) \vee (\Delta_p^{\sigma} \subseteq \Delta_q^{\sigma} \supseteq \Delta_t^{\sigma}).$$

Доказательство.

Пусть $\widetilde{R}\ (p,q)$ и $\widetilde{R}\ (q,t)$, тогда $\Delta_p^{\sigma} \cap \Delta_q^{\sigma_1} = \varnothing$, и $\Delta_q^{\sigma} \cap \Delta_t^{\sigma_1} = \varnothing$, где $\sigma \neq \sigma_1$. Если имеет место $\Delta_p^{\sigma} \subseteq \Delta_q^{\sigma} \subseteq \Delta_t^{\sigma}$ или $\Delta_p^{\sigma} \supseteq \Delta_q^{\sigma} \supseteq \Delta_t^{\sigma}$, то очевидно, что и $\Delta_p^{\sigma} \cap \Delta_t^{\sigma_1} = \varnothing$, а значит $\widetilde{R}\ (p, t)$.

Если имеет место $\Delta_p^{\sigma} \subseteq \Delta_q^{\sigma} \supseteq \Delta_t^{\sigma}$ то

$$(\Delta_p^{+} \cup \Delta_q^{+} \cup \Delta_t^{+}) \cap (\Delta_p^{-} \cup \Delta_q^{-} \cup \Delta_t^{-}) \cap$$
$$(\Delta_p^{0} \cup \Delta_q^{0} \cup \Delta_t^{0}) = \varnothing \text{ имеет место.}$$

Тогда

$$(\Delta_p^{+} \cup \Delta_t^{+}) \cap (\Delta_p^{-} \cup \Delta_t^{-}) \cap (\Delta_p^{0} \cup \Delta_t^{0}) = \varnothing.$$

А это значит, что $\widetilde{R}\ (p,t)$ имеет место и $\widetilde{R}$ – транзитивно.

Это утверждение верно для любой однородной стратегии.

Утверждение 5.

Отношение $\widetilde{R}$ для однородной стратегии с простым методом сходства транзитивно на любом упорядоченном множестве этапов.

Доказательство.

Если имеет место $\widetilde{R}\ (p,q)$ и $\widetilde{R}\ (q,t)$ и $p < q < t$, то любая гипотеза I рода, полученная на этапе p, сохраняет знак и на этапах q и t в силу непротиворечивости этих этапов. Если предположить, что $\neg\ \widetilde{R}\ (p,t)$, то значит некоторая гипотеза $J_{(1,n)}\ (V \Rightarrow_2 W)$ этапа с номером p на одном из этапов с номером m (q $< m \leq t$) стала нулевой, потому что появилась отрицательная гипотеза с таким же телом. (Других вариантов появления противоречия между этапами в однородной стратегии с простым методом сходства нет). Но в силу $\widetilde{R}\ (p,q)$ гипотеза $J_{(1,n)}(V \Rightarrow_2 W)$ сохранена на этапе с номером q и, следовательно, $\neg\ \widetilde{R}\ (q,t)$. Из того, что это противоречит условию, следует $\widetilde{R}\ (p,t)$.

Для неупорядоченных номеров этапов транзитивность может места не иметь. Если $\widetilde{R}\ (p,q)$ и

$\widetilde{R}$ (p,t), но $\neg$ $\widetilde{R}$ (q,t), то в силу симметричности отношения $\widetilde{R}$ верно и $\widetilde{R}$ (q, p). Таким образом, имеет место $\widetilde{R}$ (q, p), $\widetilde{R}$ (p,t) и $\neg$ $\widetilde{R}$ (q,t). Транзитивности в этом случае нет.

Для стратегии с простым методом сходства и запретом на контрпримеры отношение $\widetilde{R}$ транзитивно всегда, так там не возникает противоречий. А для стратегии с методом различия отношение $\widetilde{R}$ в общем случае не транзитивно в том числе и для упорядоченных номеров этапов. Возможна такая ситуация, когда на этапе с номером p есть положительная гипотеза $J_{(1,n)}$ $(V\Rightarrow_2 W)$, а на этапе с номером q эта гипотеза исчезает из-за того, что она не удовлетворяет условию d^+ метода различия. (Для остальных гипотез ничего не меняется). Тогда имеет место $\widetilde{R}$ (p,q). А на этапе с номером t из новых фактов появляется отрицательная гипотеза $J_{(-1,n)}$ $(V\Rightarrow_2 W)$. Это не нарушает непротиворечивости этапов q и t, т. е. $\widetilde{R}$ (q,t), но $\widetilde{R}$ (p,t) места иметь не будет, так как $\Delta^+_p \cap \Delta^-_t \neq \varnothing$.

Если толерантность $\widetilde{R}$ на множестве N транзитивна, но пространство $(N, \widetilde{R})$ не является полным, то это означает, что непротиворечивость множеств гипотез, получаемых в расширяющихся базах данных, нарушается, как минимум, столько раз, сколько в $(N, \widetilde{R})$ имеется классов эквивалентности.

Пусть $(N, \widetilde{R})$ имеет m классов эквивалентности. Все множества гипотез, полученные на этапах с номерами из класса эквивалентности K_i непротиворечивы между собой. Но с любым множеством гипотез, полученных на этапах с номерами из любого другого класса эквивалентности, гипотезы этапов с номерами из класса K_i вступают в противоречие. При этом противоречие может возникнуть из-за одной гипотезы, а может и из-за многих.

Поэтому самого факта наличия в пространстве $(N, \widetilde{R})$ классов эквивалентности и информации об их количестве недостаточно для того, чтобы сделать вывод о степени противоречивости ДСМ-рассуждения.

Как связаны тотальная корректность *I* и *II* родов?

Утверждение 6.

Если ДСМ-рассуждение (для любой стратегии) обладает тотальной корректностью *II* рода, то оно обладает и тотальной корректностью *I* рода.

Доказательство.

Пусть $\forall k,m$ $(1 \le k,m \le s)$ имеет место $\widetilde{K}$ (k,m).

Предположим, что $\exists p,q$ такие, что $\neg$ $\widetilde{R}$ (p, q). Пусть $\Delta^+_p \cap (\Delta^-_q \cap \Delta^0_q) \neq \varnothing$ и $J_{(1,n)}(V\Rightarrow_2 W) \in \Delta^+_p$, а $J_{(0,n)}(V\Rightarrow_2 W) \in \Delta^0_q$. Пусть также $V \subset X$, $W \subseteq Y$, и $J_{<\tau,0>}$ $(X\Rightarrow_1 Y)$ содержится в базе фактов этапа Ip и ни одна отрицательная гипотеза не включается в X. Тогда этот τ-пример будет доопределен как положительный на этапе Ip и как противоречивый на этапе Iq. Таким образом, $\neg$ $\widetilde{K}$ (p, q). Но это противоречит тотальной корректности *II* рода. Поэтому предположение о невыполнении $\widetilde{R}$ (p, q) для некоторых p и q неверно.

А вот обратное утверждение, что из тотальной корректности *I* рода следует тотальная корректность *II* рода, места не имеет.

Пусть имеет место тотальная корректность *I* рода, т. е. любая, например положительная, гипотеза *I* рода не становится ни на каком этапе отрицательной или противоречивой.

Предположим, что в базе фактов этапа Ip есть объект X, про который неизвестно, обладает или нет он свойством Y, т.е. $J_{<\tau,0>}$ $(X\Rightarrow_1 Y)$. Предположим также, что на этом этапе получена положительная гипотеза $J_{(1,n)}(V\Rightarrow_2 W)$ и $V \subset X$, $W \subseteq Y$. Если X не включает ни одну отрицательную или нулевую гипотезу, то мы доопределяем пример $J_{<\tau,,0>}$ $(X\Rightarrow_1 Y)$ как положительный.

При переходе к этапу Iq $(p < q)$ отрицательной или нулевой гипотезы с телом $(V\Rightarrow_2 W)$ не возникнет в силу $\widetilde{R}$ (p, q), но может появиться отрицательная гипотеза $J_{(-1,n)}(V_1\Rightarrow_2 W)$, $V_1 \neq V$, и тогда $J_{<\tau,0>}$ $(X\Rightarrow_1 Y)$ доопределится как $J_{(0,n)}$ $(X\Rightarrow_1 Y)$ и $\widetilde{K}$ (p, q) места иметь не будет.

Вообще, если ДСМ-рассуждение проводится с помощью рассмотренных выше однородных стратегий, то ни для одной из них тотальная корректность *II* рода не следует из выполнения ее для любой другой.

Действительно, пусть $Str_{a,n;a,n}$ порождает ДСМ-рассуждение с тотальной корректностью *II* рода. Предположим, что на этапе *I* с номером p пример $J_{<\tau,0>}$ $(X\Rightarrow_1 Y)$ доопределился как противоречивый, потому что в X содержится как V, так и V_1, такие, что $J_{(1,n)}(V\Rightarrow_2 W)$, $J_{(-1,n)}(V_1\Rightarrow_2 W)$ и $W \subseteq Y$.

Если ДСМ-рассуждение проводится с помощью стратегии $Str_{ab,n;,ab,n}$, то нулевая гипотеза *II* рода $J_{(0,n)}(X\Rightarrow_1 Y)$ этапа с номером p может стать отрицательной/положительной на этапе с номером q $(p<q)$, если появится контрпример для положительной/отрицательной причины, входящей в X.

Для $Str_{ad,n;,ad,n}$ возможна такая ситуация. Пусть в ДСМ-рассуждении для стратегии с предикатами простого положительного и отрицательного сходства на этапе с номером p появилась противоречивая гипотеза *II* рода $J_{(0,n)}$ $(X\Rightarrow_1 Y)$, содержащая положительную и отрицательную причины V и V_1. На этапе с номером q $(p<q)$ эта гипотеза сохранится в силу тотальной корректности ДСМ-рассуждения для этой стратегии. Пусть для стратегии $Str_{ad,n;,ad,n}$ на этапе с номером p пример $J_{<\tau,,0>}$ $(X\Rightarrow_1 Y)$ тоже доопределился как противоречивый. Но при расширении базы фактов на этапе с номером q $(p<q)$ условие d^- может быть не выполнено для отрицательной причины V_1, входящей в X, потому что появился пример $J_{<-1,0>}$ $(Z\Rightarrow_1 U)$ и $(X\setminus V_1) \subset Z$ и $W \subset U$ и пример $J_{(\tau,0)}$ $(X\Rightarrow_1 Y)$ тогда доопределится как положительный. Таким образом, из тотальной корректности *II* рода для стратегии с простыми предикатами сходства может не последовать тотальная корректность *II* рода для стратегий с запретами на контрпримеры и с предикатами различия. Естественно, что это же относится и к стратегии $Str_{abd,n;abd,n}$.

Нет следования и в обратную сторону. Действительно, гипотеза *I* рода $J_{(1,n)}(V \Rightarrow_2 W)$, полученная с помощью $Str_{abd,n;abd,n}$ на этапе с номером *p* такая, что $V \subset X$, $W \subseteq Y$, доопределяет пример $J_{(\tau,0)}$ $(X \Rightarrow_1 Y)$ как положительный. На этапе с номером *q* ($p<q$) из новых фактов может появиться пара (V,W), удовлетворяющая $M^-_{a,n}$, но не удовлетворяющая условиям *b* или *d*. Тогда пример $J_{<\tau,0>}$ $(X \Rightarrow_1 Y)$ на этапе с номером *q* тоже доопределится как положительный и корректность для этой стратегии не нарушится. Но при использовании стратегий $Str_{ab,n;ab,n}$ или $Str_{ad,n;ad,n}$ в силу отсутствия условий *d* или *b* соответственно, пара (V,W) уже будет удовлетворять $M^-_{ab,n}/M^-_{ad,n}$ и $J_{<\tau,0>}$ $(X \Rightarrow_1 Y)$ доопределится как противоречивый. Корректность для этих стратегий нарушится. Аналогично показывается отсутствие следования тотальной корректности *II* рода из $Str_{ab,n;ab,n}$ и $Str_{ad,n;ad,n}$ для $Str_{a,n;a,n}$.

Из того факта, что для однородных стратегий нет следования тотальной корректности *II* рода, вытекает, что они не образуют экстенсиональной решетки по вложению множеств гипотез *II* рода.

СПИСОК ЛИТЕРАТУРЫ

1. Финн В.К. Эпистемологические основания ДСМ-метода автоматического порождения гипотез. Часть I. // Научно-техническая информация. Сер. 2. – 2013. – № 9. – С. 1–29.
2. Арский Ю.М., Финн В.К. Принципы конструирования интеллектуальных систем // Информационные технологии и вычислительные системы. – 2008. – № 4. – С. 4–37.
3. Финн В.К. Индуктивные методы Д.С. Милля в системах искусственного интеллекта. Часть II. // Искусственный интеллект и принятие решений. – 2010. – № 4. – С. 14–40.
4. Шрейдер Ю.А. Равенство, сходство, порядок. – М.: Наука, 1971.
5. Гусакова С.М. Отношение толерантности, характеризующее корректность ДСМ-рассуждений // Труды XIII национальной конференции по искусственному интеллекту с международным участием – КИИ-2012 (16-20 октября 2012 г.). – Белгород, 2012. – С.100–107.
6. Гусакова С.М., Финн В. К. Сходство и правдоподобный вывод // Известия АН СССР. Сер. Техническая кибернетика. – 1987. – № 5. – С. 42– 63.
7. Финн В.К. Дистрибутивные решетки индуктивных ДСМ-процедур // Научно-техническая информация. Сер.2. – 2014. – № 11. – С. 1–29.

Материал поступил в редакцию 17.03.15.

Сведения об авторе

ГУСАКОВА Светлана Марковна – кандидат физико-математических наук, старший научный сотрудник сектора интеллектуальных информационных систем ВИНИТИ РАН.
e-mail: smg@viniti.ru

ИНФОРМАЦИОННЫЕ ЯЗЫКИ

УДК 004.89 : 54-3

Д.Л. Ольшанский, Д.А. Добрынин, В.Г. Блинова

Автоматическое кодирование химических соединений фрагментарным кодом суперпозиций подструктур*

Рассматривается проблемно-ориентированный язык описания химических соединений – фрагментарный код суперпозиции подструктур (ФКСП), его применение в исследовании связи «структура-активность». Приводится задача автоматического кодирования соединений в код ФКСП. Описываются предшествующие решения и необходимость нового кодировщика. Даются правила для ручного кодирования соединений, а также их формализация в виде решения ряда задач на графах. Рассматривается применение кодировщика в качестве компонента в составе новых интеллектуальных ДСМ-систем.

Ключевые слова: *интеллектуальная ДСМ-система, язык представления данных (ФКСП), алгоритмы на графах*

ВВЕДЕНИЕ

Важнейшей задачей в медицине, фармакологии и ряде прикладных областей органической химии является установление связей между структурой химических соединений и их биологической активностью. Решение этой задачи на ЭВМ, в связи с ее трудоемкостью, является областью интенсивных исследований.

В первую очередь, машинное решение опирается на структурные описания соединения, которые могут быть выполнены с различной степенью детализации. Таким описанием, в частности, является пространственная структурная химическая формула, которая достаточно точно представляет топологию и свойства молекулы, а также многочисленные дескрипторные языки, описывающие отдельные фрагменты соединений или функциональные группы.

Задача описания структуры химического соединения для определения корреляции «структура-активность» требует решения проблемы поиска изоморфизма подграфа и нахождения максимального общего подграфа. В общем случае – это NP-полные задачи, поэтому из соображений эффективности и простоты дальнейшей обработки веществ большинство распространенных методов полагаются на те или иные дескрипторные языки [1].

Среди различных дескрипторных языков код ФКСП (фрагментарный код суперпозиций подструктур) хорошо зарекомендовал себя для применения в компьютерных экспертных системах типа ДСМ [2, 3]. Разработанный в начале 70-х годов В.В. Авидоном, код предназначен для дискретного описания химического соединения в виде набора всех входящих в него подструктур [4]. Такими подструктурами являются центры локализации π-электронов: гетероатомы, ароматические циклические системы, комплексы кратных связей углеродов (-C=C-, C-C≡C-), соединенные цепью атомов углерода.

Теоретическая основа этого кода состоит в существовании языка распознавания структуры вещества в организме человека специальными приемниками информации – молекулярными рецепторами [2, 3]. Рецептор распознает структуру посредством образования относительно слабых химических связей с какой-то частью молекулы. Код ФКСП представляет собой семизначное число, которое содержит номера двух центров и число атомов углерода между ними, обозначающее расстояние этих центров друг от друга, а также сопряжение между ними.

Преимуществами ФКСП является хорошая структурированность данных, являющаяся необходимым требованием для работы ДСМ-системы [5], простота реализации операций пересечения, объединения и вычитания над объектами, а также получение результатов данных с меньшей долей шума, по сравнениию с работой над непосредственно структурными формулами.

К недостаткам языка ФКСП относится рассыпание структуры на фрагменты и потеря информации о связи между отдельными фрагментами-дескрипторами. Как следствие, становится невозможным выявить фармакофоры в виде более крупных связных фрагментов, также следует отметить громоздкость и сложность правил построения кода и формализации этих правил для автоматического кодирования.

Для получения существенных результатов решателю ДСМ требуется относительно большой массив

* Работа выполнена при частичной поддержке РФФИ (проект №14-07-00856а) и программы фундаментальных исследований Президиума РАН (П15, проект №209)

данных (порядка 100 соединений), а для проведения большего числа экспериментов (и установления закономерностей) требуется БД, содержащая несколько тысяч соединений. В связи с этим впервые возникла задача автоматического перевода соединений в наборы кодов ФКСП.

Первая программа для автоматического кодирования создана Лейбовым [2], в ходе этой работы был усовершенствован и дополнен язык ФКСП (ФКСП-1а). В частности были устранены некоторые нестрогости и неоднозначности языка, изначально рассчитанного на ручное кодирование. Из-за ограниченности вычислительной мощности компьютеров того времени большая часть параметров кодирования была фиксирована – таблицы и коды дескрипторных центров, метод кодирования циклических фрагментов молекул и т.д. Также неудачным оказалось применение собственного формата данных для хранения молекул.

По мере развития ДСМ-метода и расширения области его применения в фармакологии и медицине, возникла необходимость внесения изменений в кодирование химических соединений. Для исследований токсичности, канцерогенности и биотрансформации лекарственных веществ потребовалось изменить язык, дополнив его новыми центрами.

В реализации языка ФКСП-1а при кодировании циклической части молекулы выделялись все базисные несопряженные циклические структуры, ароматические группы циклов, списочные структуры, заданные таблично, и на каждую из найденных циклических структур записывался один дескриптор. При этом ароматические системы, которые состоят из множества циклов, рассматриваются как неделимые. Такой подход был обоснован тем, что при взаимодействии с рецептором, циклическая ароматическая система выступает как единый объект, в котором электронные оболочки атомов формируют единое электронное облако. В то же время, это приводило к тому, что похожие друг на друга системы циклов, отличающиеся хоть одним циклом, при пересечении не имели общих частей, поскольку дескрипторы описывали всю циклическую систему в целом. Как следствие, невозможно было выделить общие части двух циклических фрагментов молекул. Этот недостаток приобрел значение при кодировании классов соединений с преобладанием циклических систем.

Чтобы решить проблемы кодирования сложных полициклических соединений и снять ряд ограничений, в 1999 г. был создан новый кодировщик [3]. В качестве расширения языка было реализовано разбиение полициклов на цепочки, покрывающие все возможные комбинации отдельных циклов. Например, циклическая система, состоящая из четырех циклов, может включать в себя до четырех цепочек из двух циклов и три цепочки по три цикла. На каждую цепочку записывался отдельный дескриптор. Из других отличительных особенностей, в сравнении с программой Лейбова, необходимо отметить следующее:

- загрузка данных из стандартного для химических редакторов формата MDL (MOL-файлы). Это позволило легко использовать существующие базы данных по химическим соединениям;
- появилась параметризация языка – что позволило вносить коррективы в список активных центров для кодирования, а также создать подклассы языка ФКСП, адаптированные для конкретных задач;
- при кодировании сохранялась информация, какие атомы в каждой молекуле образуют тот или иной дескриптор. Это свойство использовалось средствами визуализации для изучения фрагментов соединений экспертами;
- кодировщик был реализован как динамически загружаемый модуль (DLL), что потенциально позволяло повторно использовать его в других программах для ОС семейства Windows.

Впоследствии кодировщик был интегрирован в состав интеллектуальной ДСМ-системы, которая продолжает успешно применяться в различных экспериментах [6]. Несмотря на имеющийся интерес к применению различных решателей в задачах «структура-активность», система оказалась достаточно монолитной и круг доступных в ней методов с годами не расширялся. Также эта система не была адаптирована для автоматической выгрузки результатов, в том числе для хранения и анализа экспериментов в базе знаний. При рассмотрении применения кодировщика в сетевых интеллектуальных системах с веб-интерфейсом, было обнаружено, что существенным недостатком является отсутствие поддержки ОС Linux и пакетного режима работы.

Назрела необходимость в легко отчуждаемом кодировщике, реализованном как отдельная программа, для работы в автоматическом режиме, без привязки к какому-либо пользовательскому интерфейсу. При этом необходимо было воспроизвести все правила кодирования и расширения языка ФКСП для полициклов. Также важно было сохранить основные достоинства – поддержку формата молекул и возможность настройки параметров языка. Дополнительным требованием стала возможность работы кодировщика под ОС Linux, типичной для современных серверов.

ПРАВИЛА КОДИРОВАНИЯ ФРАГМЕНТАРНЫМ КОДОМ СУПЕРПОЗИЦИЙ ПОДСТРУКТУР

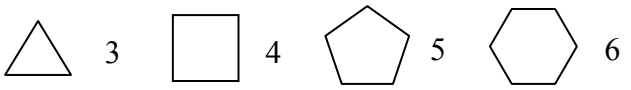
Рассмотрим все правила, применявшиеся ранее при формировании кода ФКСП и последние расширения языка [3].

В ФКСП существует три вида слов: линейный дескриптор, циклический дескриптор и дескриптор замещения.

Циклический дескриптор, описывающий одну из циклических систем в молекуле, состоит из трех частей:

«Голова»	«Ядро»	«Хвост»
Топология циклической системы	Количество π-электронов в циклически сопряженной системе	Расположение гетероатомов в циклической системе

Первая часть циклического дескриптора («голова») описывает циклическую систему в целом. Для отдельных простых циклов «голова» состоит из цифры, указывающей количество атомов в цикле.

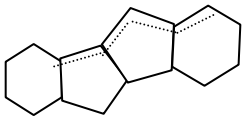


Для полициклов используется следующая запись: $x_1x_2y_1x_3y_2x_4\ldots x_n$, где x_i – размер циклов, y_i – характер связи x_i с x_{i+2} – циклом в цепочке. Характер связи записывается одной буквой (A-Z), соответственно количеству ребер (A – одно, B – два и т.д.), которым разделяются эти циклы на общей огибающей полицикла (под общей огибающей понимается внешний контур циклической системы). Для бициклических систем в запись «головы» войдут только первые две цифры.

Чтобы сохранить однозначность при кодировании, соблюдается принцип каноничности записи – из всех вариантов записи верной считается лексикографически наименьшая. Для полициклических систем соблюдается правило «непрерывной огибающей» – расстояние считается по одному краю огибающей, без прохождения через общие ребра.

Например:

Правильный код
65A5B6
Неверные коды
65A5A6, 65B5A6



Вторая часть циклического дескриптора («ядро) указывает количество π-электронов в ароматической системе. Формат записи – двузначное число, отделяющееся от «головы» запятой. Ароматической считается группа циклов, в которой число π-электронов удовлетворяет правилу Хюккеля: 4N+2, где N – натуральное число. Если система не является ароматической, то записываются цифры “00”.

Третья часть циклического дескриптора («хвост») описывает расположение гетероатомов в системе и имеет вид: $A_1B_1\ldots A_nB_n$, где A_i – символ элемента, B_i – порядковый номер вершины (правило нумерации см. ниже). Особым образом записываются символы следующих гетероатомов:

B (бор)	B
-N= (пиридиновый азот)	M
O (кислород)	Q
-N- (пирольный азот)	N
O+ (циклооксониевый кислород)	R
P (фосфор)	P
-S- (сера тиофеновая)	S
-S= (циклотиеновая сера)	T

Все остальные элементы обозначаются своими символами. В случае локализации заряда на гетероатоме, он указывается после символа, например: M+ или N+. После символа гетероатома указывается номер вершины, на которой он расположен.

Принцип нумерации вершин в циклических системах следующий:

Нумерация всегда начинается с наибольшего из крайних циклов, с первого неключевого атома. Ключевым атомом считается атом, принадлежащий нескольким циклам.

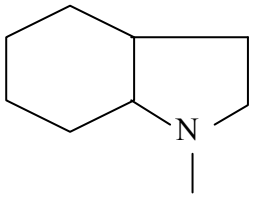
Порядок следования выбирается такой, чтобы ключевые атомы имели наибольший возможный номер.

При наличии нескольких вариантов, верным считается тот, в котором гетероатомы, стоящие вначале, имеют наименьшие номера положения, т. е. каноническая запись будет лексикографически наименьшей.

Например:

Верный код
56,00N6

Неверный код
56,00N8 56,00N1



Также к циклическим дескрипторам относятся иррегулярные (списочные) структуры (см. табл. 1), нумерация в таких структурах задается таблично. Запись при этом отличается следующим: в качестве «головы» дескриптора указывается индекс структуры, а «ядро» замещается цифрами «00» [18].

Линейный дескриптор ФКСП кодируется семизначным числом и имеет вид:

Код первого дескрипторного центра	Длина цепочки углерода	Код второго дескрипторного центра	Признак сопряжения: 0 или 1

Двузначными числами записываются дескрипторные центры и длина углеродной цепочки, а признак сопряжения – одной (1 – есть сопряжение, 0 – нет сопряжения). Дескрипторным центром является либо одиночный атом с заданными связями (см. табл. 2), либо фрагмент, состоящий из нескольких атомов (см. табл. 3). Помимо табличных в дескрипторы выделяются следующие типы центров:

ДЦ #33 – любой атом в ароматическом цикле и в циклах, где нет гетероатомов;

ДЦ #34 – любой гетероатом в ароматическом цикле;

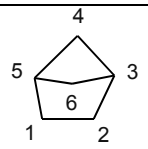
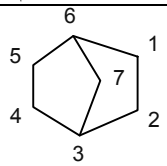
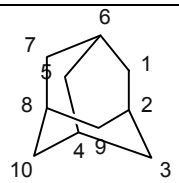
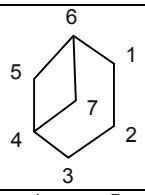
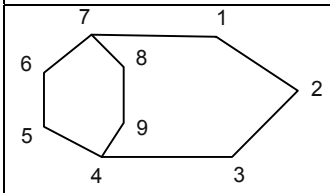
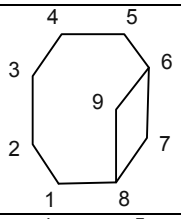
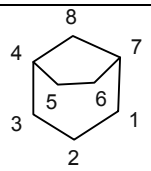
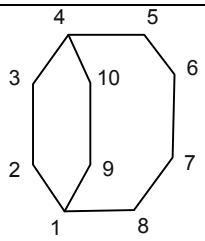
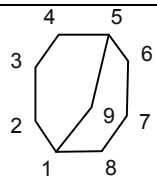
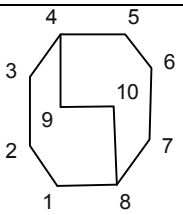
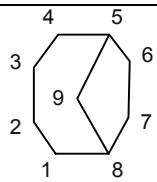
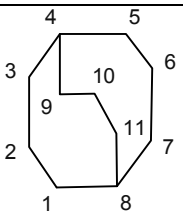
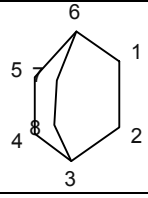
ДЦ #35 – атом углерода в ароматическом цикле, отделяемый от гетероатома в этом цикле одним ребром;

ДЦ #36 – атом углерода в ароматическом цикле, отделяемый от гетероатома в этом цикле двумя ребрами;

ДЦ #37 – атом углерода в ароматическом цикле, отделяемый от гетероатома в этом цикле тремя ребрами;

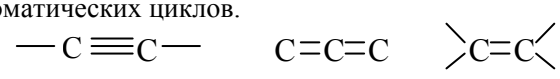
ДЦ #40 – любой заряженный гетероатом в ароматическом цикле;

Таблица 1

Структура с нумерацией атомов	Индекс структуры	Структура с нумерацией атомов	Индекс структуры
	A3		A11
	A4		A12
	A6		A13
	A7		A14
	A8		A15
	A9		A16
	A10		

ДЦ #41 – любой атом в карбоцикле (цикл состоящий только из атомов углерода в состоянии SP3-гибридизации) Также при составлении линейного дескриптора с участием этого центра, к длине углеродной цепи прибавляется N/2+1, где N- число углеродов в карбоцикле;

ДЦ #45 и #46 – углероды с кратной связью, вне ароматических циклов.



Цепочки углеродов, связывающие два ДЦ с участием 45-го или 46-го номера, не могут проходить через другие атомы с этими дескрипторами.

Когда все дескрипторные центры идентифицированы, кодируются линейные дескрипторы. Для каждой пары ДЦ проверяется наличие пути между ними – цепочки углеродов, не входящей в ароматический цикл. При этом пары 41-41, 41-42, 42-42 пропускаются. Если одним из ДЦ является центр #41 (но не карбоцикл), то к длине цепи прибавляется 1.

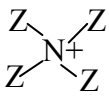
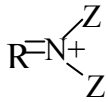
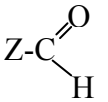
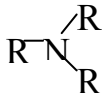
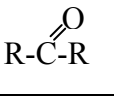
Таблица 2

Список дескрипторных центров ФКСП первого типа

атом	валентности	номер ДЦ	атом	валентности	номер ДЦ
Li	1	43	Ga	3	43
Be	1	43	Ge	4	43
B	3	53	As	3,5	51
N-	2	00	As+	4	51
O-	1	15	Se	2,4,6	54
O+	3	16	Br	1	31
F	1	32	Br	1	48
Na	1	43	Rb	1	43
Mg	2	43	Sr	2	43
Al	3	43	Y	3	43
Si	2,4	52	Zr	4	43
P	3,4,5	47	Nb	2,5	43
P+	4	47	Mo	2,4,6	43
S	6	23	Ag	1,2	43
S+	3,4	23	Cd	2	43
Cl	1	31	Sn	2,4	43
K	1	43	Sb	3,5	51
Ca	2	43	Te	2,4,6	54
Sc	3	43	I	1	31
Ti	4	43	I	1	49
V	2,3,4,5	43	Ba	2	43
Cr	2,3,4,6	43	Pt	2	43
Mn	2,4,7	43	Au	1,2	43
Fe	2,3	43	Hg	1,2	43
Co	2	43	Ti	3	43
Ni	2	43	Pb	2,4	43
Cu	1,2	43	Bi	2,3,5	43
Zn	2	43			

Таблица 3

Список дескрипторных центров ФКСП второго типа. Z – любой атом, R – любой атом, кроме H

ДЦ	валентность	код	ДЦ	валентность	код	ДЦ	валентность	код
	4	01	O=R	2	13	R-OH	2	11
	4	07	Z-SH	2	21	R-O-R	2	12
Z-NH-Z	3	02	R-S-R	2	22		2	14
	3	03	S=R	2	25	Z-CH_3	4	41
R=NH	3	04	R ≡ N	3	06	R ≡ CH	4	41, 80
R=N-R	3	05	R=CH_2	4	41, 80		4	13

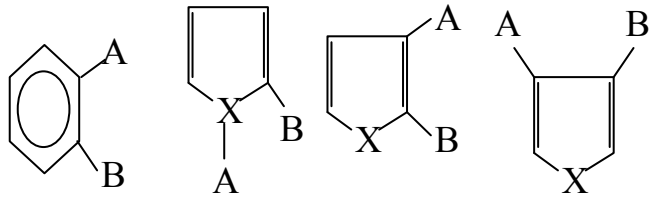
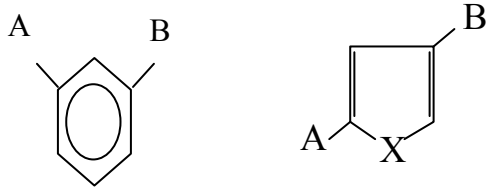
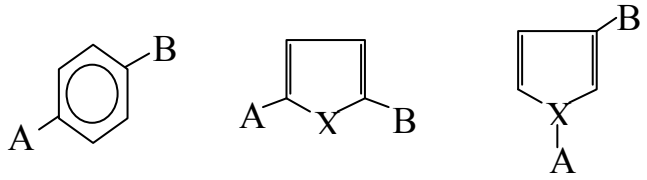
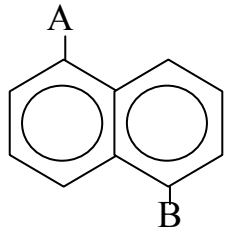
На каждую из найденных пар формируются линейные дескрипторы. Седьмой цифрой линейного дескриптора указывается цифра 0, если сопряжение отсутствует, и 1 – если сопряжение существует. Признаком сопряжения считается наличие у каждого углерода в цепи хотя бы одного π-электрона. Если присутствуют центры #41 или #42, то сопряжение проверяется только по углероду, ближайшему к ДЦ, противоположному дескрипторным центрам #41, #42.

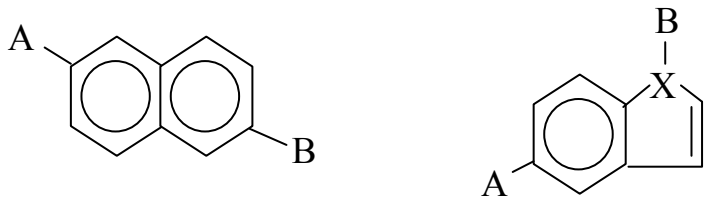
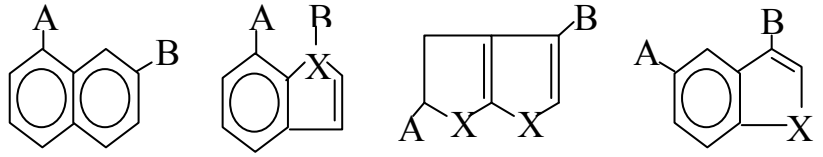
При записи линейных дескрипторов соблюдается правило каноничности – номер первого дескрипторного центра меньше, либо равен номеру второго.

Дескрипторы замещения кодируют взаимное расположение заместителей в ароматических системах и аналогично линейным дескрипторам кодируются семью цифрами. При этом так же двумя двузначными цифрами кодируются номера ДЦ заместителей, но код цепи указывается в соответствии с табл. 4. Код сопряжения (седьмая цифра) указывает возможность резонансного взаимодействия заместителей. В соответствии с теми же правилами каноничности, номер первого дескриптора должен быть меньше, либо равен номеру второго.

Таблица 4

Таблица расположения заместителей

Расположение	Код цепи	Код сопряжения
 (“орто”)	62	1
 (“мета”)	63	0
 (“пара”)	64	1
 (“пери”)	65	0
 (“амфи”)	66	1

Расположение	Код цепи	Код сопряжения
	67	1
	61	1

Перед кодированием дескрипторов замещения дополнительно выделяются дескрипторные центры #44. Этими центрами являются неароматические атомы углерода, которые находятся в состоянии SP1 или SP2 гибридизации и связаны кратной связью с каким-либо неароматическим гетероатомом.

ФОРМАЛИЗАЦИЯ ПРОЦЕССА КОДИРОВАНИЯ

Как можно заметить, процесс кодирования соединения описывается в высокоуровневых понятиях, доступных специалисту (химику), которые не имеют простых аналогов в компьютерной обработке. В действительности, все шаги процесса кодировки ФКСП можно формализовать как решение ряда задач на графах.

Граф – один из наиболее гибких абстрактных типов данных в компьютерных системах. Математически граф задается парой множеств (V, E $\in$ [VxV]), где V – множество вершин, E – множество связей между вершинами или ребер. Граф считается неориентированным, если у ребер нет ориентации т.е. для любых a,b $\in$ V пары (a,b) и (b,a) в множестве ребер считаются эквивалентными.

На практике вершины идентифицируются целым числом (индексом), а ребра – парами чисел – вершин. С вершинами и ребрами также удобно связывать некоторые значения. Естественным для химических соединений при компьютерной обработке является неориентированный граф, где атомы соответствуют вершинам, а связям – ребра, а вес ребра – кратности связи (см. рис. 1). Помимо номера из периодической таблицы, с каждым атомом необходимо связать дополнительную информация, такую как номера ДЦ, приходящиеся на данный атом. Для шаблонных фрагментов, встречающихся в описании алгоритма ФКСП, используются специальные «фальшивые» атомы, описывающие такие подстановки как «любой атом, кроме водорода» или же «любой атом».

Используя такое представление, мы формализуем основные алгоритмы кодировки ФКСП, сведя их к известным задачам из теории графов. Далее мы отметим отдельные особенности алгоритмов применительно к химическим графам и используем их ограничения для дальнейших алгоритмических усовершенствований.

Общая схема работы программы:

1. Загрузка молекулы из MOL файла(ов) во внутреннее представление. Для каждого файла выполняются шаги 2-9.
2. Разбиение молекулы на связные компоненты.
3. Поиск циклической базы и полициклических групп.
4. Кодирование ДЦ, относящихся к циклам.
5. Кодирование прочих ДЦ (1-го и 2-го рода).
6. Кодирование циклических дескрипторов.
7. Кодирование линейных дескрипторов.
8. Поиск шаблонов и кодирование дескрипторов заместителей.
9. Вывод полученных кодов.

Первый шаг кодирования – предобработка или разбиение входного файла на отдельные молекулы, поскольку стандартный формат MDL (MOL-файл) допускает наличие нескольких соединений в одном файле. Далее каждый из фрагментов можно закодировать отдельно, если же такое кодирование не желательно, то наличие нескольких молекул в одном файле свидетельствует об ошибке и также должно отслеживается. На языке теории графов – это задача поиска связанных компонентов в графе. Каждый связанный компонент и будет являться отдельной молекулой. Мы используем классическое решение посредством поиска в глубину, начиная с произвольной вершины, и маркируя все найденные в процессе этого поиска вершины. Промаркированные таким образом вершины входят в одну связанную компоненту, а поиск повторяется, начиная со следующей, еще немаркированной вершины, если таковая имеется. Подобные алгоритмы широко известны, их полное описание можно найти, например, в [7].

Поиск ДЦ первого рода достаточно прост и полностью решается для каждой вершины в изоляции (по соответствию кода элемента), при этом валентность вычисляется как сумма кратностей связей этой вершины. Поиск ДЦ второго рода решается аналогично, каждый дескриптор представляется как шаб-

лон: дескрипторный центр – атом является центром и списком соседей. Для каждой потенциальной вершины v и центра каждого шаблона t, дополнительно требуется найти соответствие между подмножеством соседей v и множеством соседей t. Из важных особенностей этой стадии алгоритма отметим, что один атом может содержать несколько ДЦ одновременно и, следовательно, каждому атому соответствует множество (обычно пустое) из номеров ДЦ.

Для дальнейшего кодирования и получения циклических ДЦ требуется идентификация всех химических циклов, и распознание их как ароматических или неароматических. В теории графов любая циклическая часть графа описывается набором циклов (базисом), из которого можно получить любой другой цикл, пользуясь операциями исключающего ИЛИ над множествами ребер базисных циклов. С точки зрения химии интерес представляет базис наименьшей суммарной длины, так как именно он состоит из всех «простых» циклов. Для решения этой задачи применяется вариация алгоритма Хорстмана [8]. Алгоритм основан на теореме: каждый цикл из минимального базиса является фундаментальным циклом. Цикл является фундаментальным, если он получен из остовного дерева (рис. 2) замыканием двух кратчайших путей, не имеющих общих ребер. Для каждой вершины графа есть остовное дерево, и, таким образом, число циклов, которые необходимо проверить, сводится к O(NM), где N – число вершин, M – ребер в графе. Далее построение базиса из этих циклов производится исключением по правилу Гаусса: очередной цикл отбрасывается, если его можно представить как симметрическую разницу циклов (см. рис. 2) уже входящих в базис. Таким образом, отбирая циклы в базис, начиная с меньших, мы получим наименьший циклический базис.

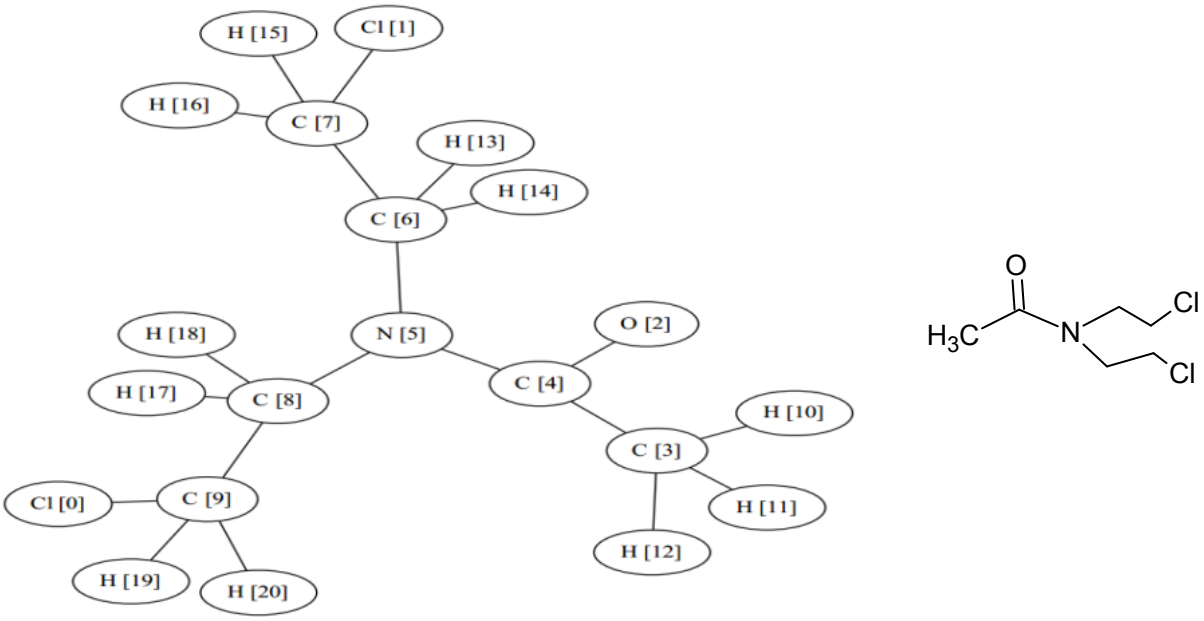


Рис. 1. Молекула и ее граф.

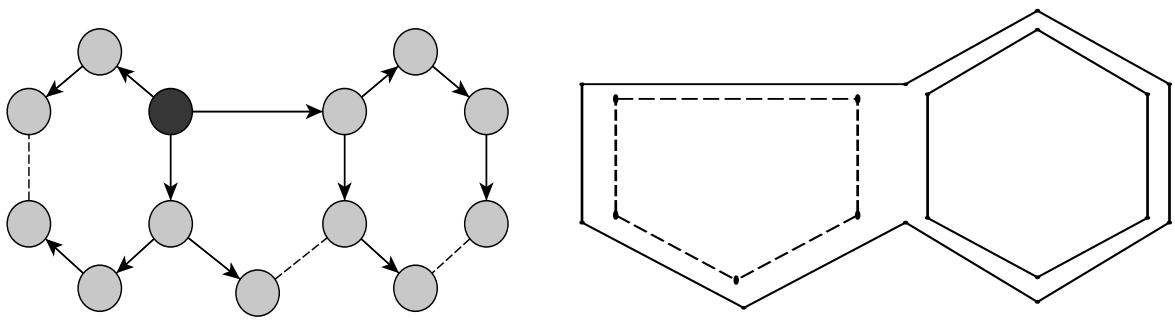


Рис. 2. Остовное дерево (слева) и симметрическая разница для циклов (справа)

Полициклические группы выделяются посредством поиска связанных компонент среди вершин, входящих в циклический базис. Для каждой связной компоненты из циклического базиса составляется граф циклов – вершины этого графа – циклы, а ребра описывают пересечения. Для кодирования всех цепочек циклов, решая задачу перебором, потребуется проверить до 2^N комбинаций, где N – полное число циклов в группе (каждый цикл может входить или не входить в цепочку). Существенно сократить переборы можно, опираясь на связность полученного полицикла, применяя следующий алгоритм. Можно начать с полного цикла (который по определению связан) и удалять из него каждый цикл из полицикла по одному. Если связность множества оставшихся циклов не теряется (т.е. мы не разбили полицикл), то кодируем эту цепочку. Повторяя данную процедуру, мы получим все цепочки, при этом отсекая целые поддеревья выбора, если множество циклов оказывается не связно. Чтобы избежать повторов комбинаций, применяем правило каноничности – отбрасывать циклы можно лишь с номерами большими последнего отброшенного. Ниже приведен псевдокод данного алгоритма. Условные обозначения: Ci – i-й цикл в группе, connected – процедура проверки связности полицикла, encode – процедура кодирования полицикла, 1...n – диапазон включающий числа от 1 до n включительно.

```
function encode_poly:
    A = C1 ∪ C2 ... ∪ Cn
    encode(A)
    for i in 1..n:
        A = A \ Ci
        if connected(A):
            split(A, i+1)
```

```
function split(A, k):
    encode(A)
    for j in k..n:
        A = A \ Cj
        if connected(A):
            split(A, j+1)
```

Понятие ароматичности циклов выполняется подсчетом числа π-электронов, заданных таблично (табл. 5) и проверкой правила Хюкеля для каждой кодируемой циклической системы. При этом алгоритм разбиения циклов можно видоизменить, запретив разбиение ароматических систем, тем самым восстановив принцип ФКСП-1а в кодировании ароматических групп.

Таблица 5

Таблица π-электронов для различных атомов по количеству кратных связей: 1- только одинарные связи связей, 2 – есть одна двойная, 3 – есть две двойных, 4 – есть одна тройная.

Имя элемента	1	2	3	4	Имя элемента	1	2	3	4
H	0	-	-	-	Cu	2	-	-	-
Li	0	-	-	-	Zn	2	-	-	-
Be	2	-	-	-	Ga	2	-	-	-
B	2	1	-	-	Ge	2	1	-	-
C	0	1	2	1	As	2	1	2	-
C	1	-	-	-	Se	2	1	2	-
C	1	-	-	-	Se	1	1	-	-
N	2	1	-	-	Br	0	-	-	-
N	1	1	1	1	Rb	2	-	-	-
N	2	1	-	-	Sr	2	-	-	-
O	2	1	-	-	Y	2	-	-	-
O	1	1	-	-	Zr	2	-	-	-
O	1	-	-	-	Nb	2	-	-	-
F	1	-	-	-	Mo	2	-	-	-
Na	0	-	-	-	Tc	2	-	-	-
Mg	2	-	-	-	Ru	2	-	-	-
Al	2	-	-	-	Rh	2	-	-	-
Si	2	1	2	-	Pb	2	-	-	-
P	2	1	2	-	Ag	2	-	-	-
P	1	-	-	-	Cd	2	-	-	-
S	2	1	2	-	In	2	-	-	-
S	1	1	-	-	Sn	2	-	-	-

Имя элемента	1	2	3	4	Имя элемента	1	2	3	4
Cl	0	-	-	-	Sb	2	-	-	-
K	0	-	-	-	Te	2	1	2	-
Ca	2	-	-	-	I	0	-	-	-
Sc	2	-	-	-	Ba	2	-	-	-
Ti	2	-	-	-	W	2	-	-	-
V	2	-	-	-	Pt	2	-	-	-
Cr	2	-	-	-	Au	2	-	-	-
Mn	2	-	-	-	Hg	2	-	-	-
Fe	2	-	-	-	Tl	2	-	-	-
Co	2	-	-	-	Bi	2	-	-	-
Ni	2	-	-	-	Bi	1	-	-	-

Для каждой выделенной циклической группы можно перейти к самому процессу кодирования циклических дескрипторов. Для записи «головы» дескриптора используется правило «непрерывной огибающей». Огибающая вычисляется как симметрическая разность множеств ребер всех входящих в группу циклов. Заметим, что огибающая обходится дважды, в одном случае для получения записи «головы», при этом выбирается наименьший крайний цикл, и во второй раз – для получения нумерации вершин, при этом нас интересует больший цикл.

Для кодирования головы вычисляется до четырех вариантов кодирования: потенциально существует до двух циклов минимальной длины, находящихся «с краю» группы, для каждого из которых есть два направления обхода: по часовой и против часовой стрелки. По правилу каноничности – верной будет лексикографически наименьшая запись из четырех возможных. Аналогичная вариативность есть и при вычислении порядка атомов в цикле для записи «хвоста» дескриптора, но уже до восьми комбинаций (рис. 3): до двух наибольших циклов «с краю» группы, имеется один или два первых атома, не принадлежащих другому циклу (неключевых) и два направления обхода.

Кодирование линейных дескрипторов требует найти (кратчайший) путь по атомам углерода между каждой парой ДЦ, если таковой имеется, также непроходимыми считаются атомы, входящие в ароматический цикл. Все уникальные пары можно получить предварительно отсортировав список ДЦ, и используя правило каноничности записи: первый ДЦ в записи должен быть меньше, либо равен второму. Поскольку количество пар невелико, для поиска кратчайшего пути мы используем алгоритм Дикстры [9] для каждой пары вершин, опять же имея в виду, что в одной вершине может быть несколько ДЦ. В качестве оптимизации, вершины графа с атомами водорода при поиске пути игнорируются, мы используем знания предметной области для упрощения задачи и в дальнейшем (рис. 4).

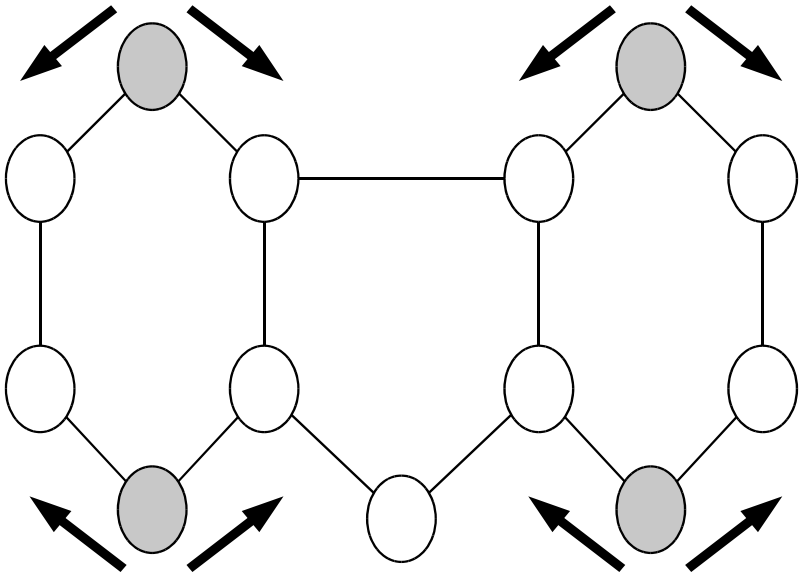


Рис. 3. Максимальное число вариантов обхода полицикла при записи хвоста

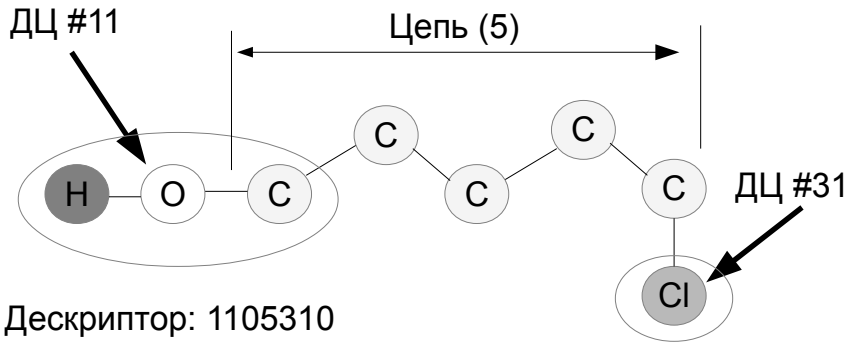


Рис. 4. Кодирование линейных дескрипторов

Задача поиска списочных структур и дескрипторов замещения сводится к поиску подграфа, изоморфного некоторому графу из списка образцов. Изоморфизм есть однозначное отображение вершин и ребер одного графа другому, при этом веса (тип атомов и кратности) также должны соответствовать друг другу. В случае дескрипторов замещения образец параметризован, в том числе по наличию двух ДЦ на определенных позициях в искомом подграфе. Список образцов задается на этапе инициализации и является одним из внешних параметров задачи.

Эффективный алгоритм поиска изоморфизмов на графах, в общем случае, одна из нерешенных проблем в теории графов. Для текущей задачи есть ряд ограничений, упрощающих решение, так что теоретически худшие случаи на практике не реализуются. Ребра описываются небольшим числом (1-3); атомы в молекуле расположены характерно, например водород, встречается лишь на тупиковых ветвях графа; атомы с ДЦ достаточно редки и часто уникальны, что радикально сокращает количество вариантов. Мы используем реализацию одного из первых алгоритмов, решающих эту задачу, алгоритм Ульмана [10].

ТЕСТИРОВАНИЕ

Для проверки работоспособности нового кодировщика проводился ряд испытаний по различным массивам соединений, среди которых есть как показательные примеры (выборка FCSS), так и реальные данные прошлых экспериментов (табл. 6). В отдельных случаях работа кодировщика проверялась экспертом, для остальных выполнялось сравнение с результатами, полученными прежним кодировщиком.

Также указаны предварительные результаты производительности, время кодирования измерялось программой time из пакета coreutil Linux, выбиралось лучшее из трех запусков. Измерения проводились на ноутбуке с процессором Intel i5-3317U 1.7 ГГц. Эта предварительная оценка показывает, что, похоже, именно кодирование сложных циклов является наиболее трудоемким процессом.

Таблица 6

Список тестовых массивов соединений

Название	Размер и характер соединений	Полное время кодирования
FCSS	41 (показательные примеры)	0.08 сек
Polycyclic	90 (преобладание циклов)	0.65 сек
Liverpul	267 (в основном простые циклы)	0.78 сек
Halogen	63 (линейные молекулы)	0.09 сек

При тестировании также были выявлены некоторые расхождения с прежним кодировщиком при записи «хвоста» циклических дескрипторов в нумерации атомов. Нумерация в новом кодировщике следует правилам ФКСП-1а, однако, возникают сомнения в следовании этим правилам нумерации прежним кодировщиком. В то же время, отличная от описанной ранее нумерация вершин не могла оказать существенного влияния на работу ДСМ-систем, поскольку единый принцип использовался для всех соединений, что никак не отразилось на поиске сходства в структуре.

ЗАКЛЮЧЕНИЕ

В отличие от предшествующих реализаций [5, 6], кодировщик реализован на широко распространенном языке C++, с применением лишь стандартных и открытых библиотек Boost C++. В частности, используется библиотека для обработки графов Boost Graph. В будущем это позволит осуществить сборку программы практически под любую операционную систему и облегчит использование вместе с существующими системами. Также существенно упрощен механизм взаимодействия – программа работает в пакетном режиме: в командной строке задается набор входных файлов и параметров, а результат выдается на стандартный вывод.

Одним из ближайших применений нового кодировщика будет проект сетевой среды для интеллектуальных ДСМ-систем. Этот комплекс объединит в себе несколько существующих решателей, базу знаний и пользовательский веб-интерфейс. Испытание покажет, насколько успешно новый кодировщик может работать совместно с другими компонентами системы.

СПИСОК ЛИТЕРАТУРЫ

1. Джурс П., Айзенауэр Т. Распознавание образов в химии. – М.: Мир, 1979. – 230 с.
2. Лейбов А.Е. Автоматическое кодирование химических структур кодом ФКСП // Итоги науки и техники. Сер. «Информатика». – 1991. – Т.15. – С.141-158.
3. Блинова В.Г., Добрынин Д.А. Языки представления химических структур в интеллектуальных системах для конструирования лекарств // Научно-техническая информация. Сер. 2. – 2000. – №6. – С. 14-21.
4. Avidon V.V., Pomerantsev A.B. Structure-activity relationship oriented languges for chemical structure representation // J. Chem. Inf. Comput. Sci. – 1982 – Vol.22, №4. – P. 207-214.
5. Финн В.К. Правдоподобные рассуждения в интеллектуальных системах типа ДСМ // Итоги науки и техники. Сер. «Информатика». – 1991. – Т. 15.– С. 54–101.
6. Блинова В.Г., Харчевникова Н.В., Добрынин Д.А., Врачко М. Интеллектуальный анализ канцерогенности химических соединений с помощью стратегий метода, основанного на логике Джона Стюарта Милля // Сборник материалов XVII Российского национального конгресса «Человек и лекарство», Москва, 12-16 апреля 2010 г. – М., 2010. – 738 с.
7. Hopcroft J., Tarjan R. Efficient algorithms for graph manipulation // Communications of the ACM. – 1973. – Vol. 16 (6). – P. 372–378.
8. Horton J. D. A polynomial-time algorithm to find the shortest cycle basis of a graph // SIAM Journal on Computing. – 1987. – B. 2, T. 16. – P. 358–366.
9. Седжвик Р. Фундаментальные алгоритмы на C++, 3-я ред. – Киев, 2011. – 483 с.
10. Ulman J.R. An algorithm for subgraph isomorphism // Journal of the ACM. – 1976. – Vol. 23. – P. 31-42.

Материал поступил в редакцию 09.03.15.

Сведения об авторах

ОЛЬШАНСКИЙ Дмитрий Леонидович – аспирант, Всероссийский институт научной и технической информации Российской академии наук
e-mail: dmitry.olsh@gmail.com

ДОБРЫНИН Дмитрий Анатольевич – кандидат технических наук, научный сотрудник, Всероссийский институт научной и технической информации Российской академии наук
e-mail: dobr@viniti.ru

БЛИНОВА Валентина Георгиевна – кандидат химических наук, старший научный сотрудник Всероссийский институт научной и технической информации Российской академии наук
e-mail: blinova@viniti.ru

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 81’322.2 : 001.102 – 027.21

В.А. Яцко

Метод автоматической классификации текстов, основанный на законе Ципфа

Описывается метод автоматической классификации текстов, основанный на анализе отклонения распределения слов от закона Ципфа в сочетании с зональной обработкой данных. Под отклонением понимается разница между реальным числовым коэффициентом слова и коэффициентом, который у него должен быть в соответствии с законом Ципфа. Применение метода предусматривает разбивку входного и эталонного текстов на зоны J_0, J_1, J_2 и создание на основе слов, входящих в зону J_0, числового ряда, в котором указываются разницы между реальными коэффициентами слов и коэффициентами, вычисляемыми по закону Ципфа. Предложенный метод позволяет существенно снизить размерность текстов и повысить быстродействие автоматической классификации.

Ключевые слова: *закон Ципфа, зональная обработка текстов, автоматическая классификация текстовых документов, повышение эффективности*

Ранее [1] нами было предложено понятие лингвистической информатики для обозначения дисциплины, изучающей закономерности распределения текстовой информации, проблемы, принципы, методы и алгоритмы разработки лингвистического программного обеспечения и аппаратных средств. Заметим, что данная интерпретация отличается от подхода, принятого в англо-американской науке, в соответствии с которым изучение прикладных проблем разработки лингвистического программного обеспечения соотносится с *компьютерной лингвистикой*, а законы и закономерности распределения текстовой информации изучаются в рамках *квантитативной лингвистики* [2]. Предложенный нами термин продолжает традиции советской науки, поскольку термин *информатика* был, как известно, предложен советскими учёными [3] для обозначения дисциплины, сочетающей как теоретические, так и прикладные аспекты.

Одним из наиболее известных и широко применяемых в предметной области лингвистической информатики законов является закон Ципфа, устанавливающий зависимость между рангом слова и его частотностью в тексте [4], которая может быть выражена формулой

$$F_r \propto 1/r^a \qquad (1),$$

где F – частотность данного слова, r – его ранг в ранжированном списке, а экспонента a примерно равна 1.

Закон Ципфа имеет предсказательную силу: зная частотность и ранг данного слова, можно определить частотности и ранги всех остальных слов в данном тексте. Если, допустим, десятое по рангу слово повторяется в данном тексте 36 раз, то частотность пятого по рангу слова будет равна 36*10/5=72.

Закон Ципфа применялся с целью сокращения размера индекса и повышения быстродействия информационно-поисковых систем [5]; стратификации текстов [6]; типологической классификации языков [7]. Выяснилось, что распределение населения по городам также соответствует этому закону [8].

В настоящей статье будет показана возможность использования закона Ципфа в сочетании с методом зональной обработки в целях автоматической классификации и авторской атрибуции текстов. В этой связи будут рассмотрены основные проблемы автоматической классификации текстовых документов, а также описан метод зональной обработки текстов.

Цель автоматической классификации текстовых документов – распознать класс, к которому относится данный текст, на основе анализа его содержания. На входе у программы-классификатора – текстовый документ, на выходе – имя класса, к которому он относится [9]. Автоматическая классификация текстов широко используется в различных лингвистических приложениях и системах, включая информационно-поисковые системы, программы фильтрации спама, электронные библиотеки, системы распознавания плагиата и авторской атрибуции. К настоящему времени сложилось два подхода к решению проблемы автоматической классификации, которые можно назвать словарным и дистантным. В основе словарного подхода лежит создание эталонного словаря класса, каждая единица которого обладает высокой дискриминирующей силой, т.е. способностью уникально

идентифицировать данный класс *C*, отличая его от другого класса/классов. В том случае, если определенный класс сопоставляется только с одним другим классом, обозначаемым как ~*C*, говорят о бинарной классификации. Типичным примером бинарной классификации является задача фильтрации спама, решение которой предусматривает отнесение входного текста либо к классу «спам», либо к классу «не спам».

Для создания эталонных словарей применяются достаточно сложные метрики, такие как хи-квадрат, отношение шансов (odds ratio), прирост информации (information gain) [10], которые позволяют выявить дискриминирующую силу лексических единиц. Наряду с эталонным словарем данного класса создаётся также отрицательный словарь, в который входят единицы, представляющие класс ~*C*. Если во входном (тестовом) документе находится слово из эталонного словаря, то ему начисляется положительный коэффициент (например, 1); словам из отрицательного словаря начисляется отрицательный коэффициент (например, -1). Далее находится сумма положительных и отрицательных коэффициентов, и, если она превышает некоторый пороговый уровень, то входной документ соотносится с классом *C*; если пороговый уровень не превышен, то текст соотносится с классом ~*C* (в случае бинарной классификации), либо игнорируется [11].

Дистантный подход основан на вычислении расстояния между параметрами входного текста и параметрами эталонного текста либо некоторой модели эталонного текста. Под параметром понимается единица текста с приписанным весовым коэффициентом. Наиболее распространённый метод вычисления расстояний между текстами предусматривает их векторное моделирование. Каждый параметр текста представляется в виде точки в многомерном пространстве, а соотношение между соответствующими параметрами во входном и эталонном текстах определяется направлением (обычно от входного к эталонному тексту) и длиной вектора, которая вычисляется на основе соотношения весовых коэффициентов. Расстояние между входным и эталонным текстами вычисляется как сумма абсолютных разниц между величинами векторов. Чем меньше расстояние, тем больше вероятность того, что входной текст относится к классу, представленному эталонным текстом. В рамках дистантного подхода также применяются пороговые уровни: решение об отнесении входного текста к классу *C* принимается, если расстояние между данным текстом и эталонным текстом меньше некоторого порогового уровня.

Существенным отличием дистантного подхода является то, что в качестве параметров могут использоваться стоп-слова, в то время как в словарном подходе производится их фильтрация. Артикли, предлоги, местоимения, союзы встречаются во всех текстах независимо от жанрово-стилистических особенностей, и различия в их распределении могут использоваться с целью вычисления расстояния между текстами.

В нашей статье будет применён дистантный подход в сочетании с методом зонально-корреляционной обработки текста, который был подробно описан ранее [12, 13]. В качестве эталонного текста нами был составлен файл, включающий пять романов Т. Драйзера (*The Genius*, *The Financier*, *The Titan*, *Sister Carrie*, *Genie Gerhardt*), которые были взяты со страницы проекта «Гуттенберг»[1]. Тексты были отредактированы, из них была удалена информация о самом проекте. В результате получился текст из 23591 уникальных слов и 1003944 общих токенов. Статистические данные о распределении токенов были получены с помощью *AntConc 3.1.3 concordancer*[2]. Выбор работ Драйзера мотивировался следующими причинами. Для художественной литературы характерно разнообразие лексики, которое не типично для других стилей. Например, в научных текстах используется более стандартизированная терминология, состав которой зависит от определённой предметной области. Работы Драйзера относятся к классической литературе, в них содержится меньше неологизмов, жаргонизмов, редких слов, чем, например, в научной фантастике. Работы автора относят к натуралистическому направлению в литературе[3], которое существенно отличает их от работ других авторов. Это позволяет получить наиболее наглядные результаты при сопоставлении текстов Драйзера с текстами других авторов. Романы автора достаточно объёмны, что позволяет составить файл с эталонным текстом размером миллион общих токенов. Именно такой размер удовлетворяет требованию репрезентативности, поскольку закон Ципфа выполняется на Брауновском корпусе, содержащем один миллион токенов [14].

В качестве тестовых документов были выбраны роман Т. Драйзера *The Stoic* и книга Ч. Диккенса *David Copperefield*. Работы Диккенса также относятся к классической литературе, однако, он был английским, а не американским писателем и жил раньше Драйзера. Это, как мы полагаем, позволит, с одной стороны, найти достаточно много общих слов и схожих параметров, а с другой стороны, наглядно продемонстрировать различия в распределении параметров. Расстояние (*Ds*) между текстом, написанным Диккенсом (файл *Di*), и эталонным текстом, включающим упомянутые работы Драйзера (файл *Dr1*), должно быть больше, чем расстояние между эталонным текстом и другой работой Драйзера, романом *The Stoic* (файл *Dr2*). Таким образом, следует найти:

$$\mathrm{Ds}(Dr1, Di) = |P(Dr1 - P(Di)| \quad (2);$$

$$\mathrm{Ds}(Dr1, Dr2) = |P(Dr1 - P(Dr2)| \quad (3),$$

где P – некоторый параметр. Очевидно, что Ds(*Dr1, Di*) должно быть больше, чем Ds(*Dr1, Dr2*). Предполагается, что использование закона Ципфа в сочетании с методом зональной обработки текстов позволит выявить существенную разницу между Ds(*Dr1, Di*) и Ds(*Dr1, Dr2*). В подтверждение данной гипотезы были выполнены следующие процедуры.

[1] https://www.gutenberg.org/. В рамках проекта проводится оцифровка и редактирование работ писателей-классиков, на которые не распространяется авторское право.

[2] http://www.laurenceanthony.net/software.html.

[3] http://www.chitai.kraslib.ru/28.html.

1. Зональная обработка текстов. Метод зональной обработки предусматривает разбивку каждого текста на три зоны: J_0, J_1, J_2. Зона J_0 включает стоп-слова и содержит наименьшее количество элементов с наибольшими частотностями, соответственно её можно назвать зоной концентрации информации. Зона J_1 состоит из знаменательных слов, отражающих содержание текста. В зоне J_2 находятся редко используемые слова, сокращения, авторские неологизмы. Это – зона наибольшего рассеивания информации, поскольку она содержит наибольшее количество элементов с наименьшими частотностями.

Разбивка текстов на зоны выполняется на основе системы уравнений:

$$\begin{cases} S(J2) = C / K \\ K = (q^n - 1) / (q - 1) \\ S(J1) = S(J2) * q \\ S(J0) = S(J1) * q \end{cases} \quad (4),$$

где C – сумма числового ряда, элементами которого являются частотности слов текста; $S(J_1)$, $S(J_2)$, $S(J_0)$ – числовые значения соответствующих зон; n = 3 – константа, равная количеству зон; q – зональный коэффициент, который находится эмпирическим путём и значение которого зависит от конкретной предметной области. Ранее на основе анализа распределения стоп-слов было показано, что для художественных текстов оптимальное значение $q = 3$ [12]. На основе формул (4) вначале находится абстрактный пороговый уровень, представляющий количественное значение каждой зоны, а затем реальный пороговый уровень, максимально близкий к абстрактному и равный сумме частотностей слов, входящих в данную зону.

В табл. 1 представлены результаты зональной обработки. В колонке «Количественные значения» указаны реальные пороговые уровни.

2. Взвешивание слов и образование параметров. В результате взвешивания каждому слову в тексте был приписан вероятностный коэффициент, который вычислялся по формуле

$$P = \frac{f(w_{ij})}{\sum_{j=1}^{n} f(w_j)}, \quad (5)$$

где f – частотность слова w_i в тексте j-м. Коэффициенты округлялись до семи десятичных знаков.

В качестве параметра было использовано среднее квадратичное отклонение от распределения Ципфа в зонах J_0 трёх текстов, которое находилось по формуле

$$\sigma(Rx) = \sqrt{Var(Rx)}, \quad (6)$$

где Var – дисперсия, а Rx – числовой ряд:

$$R(w_i...w_n) = |P(w_i...w_n) - Z(w_i...w_n)|, \quad (7)$$

где P – вероятностный коэффициент, приписанный каждому слову в зонах J_0, а Z – числовое значение слова, соответствующее закону Ципфа (распределение Ципфа). Распределение Ципфа подсчитывалось по формуле

$$P(w_{ij}) = P(w1_j) / R(w_{ij}) \quad (8)$$

где $P(w1_j)$ – вероятностный коэффициент первого по рангу слова, а R – номер ранга слова. Если $P(w1)$ = 0.0376728, а коэффициент второго по рангу слова $P(w2)$= 0.0368959, то распределение Ципфа $Z(w2)$ = 0.0188364, а отклонение от этого распределения $R(w2)$=|0.0368959-0.0188364|=0.018059. В соответствии с формулой (8) $P(w1)=Z(w1)$.

Ещё одним параметром была средняя сумма разностей в числовом ряду Rx:

$$Ms = \frac{\sum R(w_i...w_n)}{x}, \quad (9)$$

где x – количество слов в Rx.

Таблица 1

Результаты зональной обработки текстов

Текст	Зоны	Количественные значения	Диапазон слов	Кол-во слов в зоне	Кол-во общих токенов	Кол-во уникальных слов
Di	$S(J_0)$	253073	1–228	228	365542	14131
	$S(J_1)$	84359	229–2725	2497		
	$S(J_2)$	28110	2726–14131	11406		
Dr2	$S(J_0)$	88253	1–278	278	127463	9390
	$S(J_1)$	29413	279–2793	2515		
	$S(J_2)$	9797	2794–9390	6597		
Dr1	$S(J_0)$	694910	1–287	287	1003944	23591
	$S(J_1)$	231681	288–3472	3185		
	$S(J_2)$	77353	3473–23591	20119		

3. Сопоставление распределения параметров по зонам J_0 трёх текстов и вычисление расстояний между текстами. Результаты анализа распределения параметров, приведённые в табл. 2, подтверждают сформулированную ранее гипотезу о том, что применение предложенного метода позволяет установить, что расстояние между зонами J_0 текстов, написанных одним автором, существенно меньше (на 789.70% и 249.21%) , чем расстояние между этими же зонами текстов, написанных разными авторами.

Для того чтобы подтвердить эффективность использования зонального анализа в целях классификации текстов, был дополнительно проведён корреляционный анализ распределения слов по зонам J_1 трёх текстов. Как было сказано ранее, эти зоны содержат знаменательные слова, отражающие содержание текстов. Можно ожидать, что зоны J_1 в работах Драйзера $J_1(Dr1)$и $J_1(Dr2)$ содержат больше идентичных слов, чем соответствующие зоны в эталонном тексте Драйзера $J_1(Dr1)$ и тексте Дикенса $J_1(Di)$. Соответственно, сумма параметров этих зон также должна иметь большее числовое значение. С целью подтверждения этой гипотезы были проведены следующие процедуры.

1. Выполнено пересечение зон J_1 трёх текстов и найдены идентичные слова, входящие в эти зоны.

$$A=J_1(Dr1)\cap J_1(Dr2) \quad (10)$$

$$B=J_1(Dr1)\cap J_1(Di)\ . \quad (11)$$

Пересечение выполнялось с помощью стандартных функций MS Excel 2010 functions ЕСЛИОШИБКА, ВПР, ЛОЖЬ. В табл. 3 показаны результаты пересечения. Как и ожидалось, количество слов в A оказалось больше, чем в B.

2. Получена сумма вероятностных величин слов в A и B. Каждое слово в области пересечения имело два коэффициента, которые и суммировались:

$$\Sigma P(Dr1) \text{ и } \Sigma P(Dr2);\ \Sigma P(Dr1) \text{ и } \Sigma P(Di).$$

3. Вычислено среднее вероятностное значение $Mp=\dfrac{\sum P(T_i)}{x}$, где x – количество слов в данной зоне J_1.

4. Получена сумма средних вероятностных значений.

$$Mp(A)=\frac{\sum P(Dr1)}{x}+\frac{\sum P(Dr2)}{x}=0.0001228 \quad (12)$$

$$Mp(B)=\frac{\sum P(Dr1)}{x}+\frac{\sum P(Di)}{x}=0.0001145 \quad (13)$$

Таким образом, $Mp(A)$ на 7.21% больше, чем $Mp(B)$, что подтверждает сформулированную ранее гипотезу.

В настоящей статье был предложен метод классификации текстовых документов, основанный на анализе распределения Ципфа в сочетании с зональной обработкой данных. Этот метод предусматривает разбивку входных документов и эталонного текста на три зоны и сопоставление распределения слов в зонах J_0. Создаётся четыре числовых ряда, один из которых включает слова и сырые частотности; второй – слова и вероятностные величины, вычисляемые на основе частотностей; третий – слова и вероятностные величины, вычисляемые на основе закона Ципфа; четвёртый – вычисляемые по модулю разницы между величинами в третьем и втором числовом ряду. Таким образом, четвёртый числовой ряд включает величины, которые указывают на степень отклонения распределения слов в зоне J_0 от закона Ципфа. Далее сопоставляется распределение величин в четвёртом числовом ряду во входных текстах и эталонном тексте и вычисляется расстояние между ними, на основе чего принимается решение об отнесении входных текстов к классу, представленному эталонным текстом.

Одной из основных проблем, отмечаемой в работах по автоматической классификации текстовых документов [15], является их большой объём. Обрабатываемые тексты могут включать тысячи и даже десятки тысяч слов, и анализ их распределения отрицательно влияет на быстродействие системы. Предложенный подход позволяет, во-первых, существенно снизить размерность текстов, ограничив их анализ зонами J_0, которые включают несколько сотен слов, а во-вторых, упростить математический аппарат за счёт применения простых вероятностных величин.

Следует отметить, что анализ отклонения распределения слов от закона Ципфа можно применять только к зоне J_0. В остальных двух зонах многие слова имеют одну и ту же частотность и недостаточный разброс значений. В табл. 4 показано, что слова, которые встречаются в тексте один и два раза, составляют почти половину всех слов в тексте.

Таблица 2

Расстояния между эталонным текстом *Dr1* и тестовыми документами *Dr2*, *Di*, вычисленные по двум параметрам

Параметр	Dr2	Dr1	Di	Ds(Dr1,Dr2)	Ds(Dr1,Di)	Разница расстояний (%)
Ms	0.0015843	0.0015262	0.00204367	0.0000582	0.0005175	789.70%
σ(*Rx*)	0.0027441	0.0025749	0.0031659	0.0001692	0.0005910	249.21%

Таблица 3

Пересечение зон J_1 трёх текстов

Область пересечения	Количество слов	5 произвольно выбранных слов	P(Dr1)	P(Dr2)	P(Di)
A	1818	returned	0.0003277	0.0004158	
		our	0.0002341	0.0004080	
		understand	0.0002988	0.0004080	
		able	0.0002829	0.0004001	
		believe	0.0003556	0.0004001	
B	1639	master	0.0000588		0.0005307
		child	0.0002839		0.0005170
		boy	0.0002161		0.0005116
		cried	0.0000647		0.0005116
		name	0.0003217		0.0005116

Таблица 4

Частотности слов в зоне J_2 эталонного текста

Частотность слов	Диапазон	Количество слов	Процент от общего количества слов	Примеры слов
1	15535–23591	8057	34,15%	ziner zithers zouave
2	12163–15534	3372	14,29%	sapient sappho sarah
3	10179–12162	1984	8,41%	inanimate incarnation incense
4	8822–10178	1357	5,75%	tolerate tom tongues
5	7848–8821	974	4,13%	tinder tinsel touring
6	7135–7847	713	3,02%	pitied planet planted
7	6510–7134	625	2,65%	circular civic clay
8	6059–6509	451	1,91%	excuses exercised existing
9	5683–6058	376	1,59%	debt defeated defend
10	5335–5682	348	1,48%	foul freight frock

В зоне J_0 разброс частотностей намного выше, что позволяет объяснить тот факт, что значения *Ms* и $\sigma(Rx)$ на порядок выше, чем значения *Mp(A) Mp(B)*, которые вычислялись на основе распределения параметров в зоне J_1.

В целом, полученные результаты свидетельствуют, что предложенный метод может эффективно применяться в целях автоматической классификации текстовых документов.

СПИСОК ЛИТЕРАТУРЫ

1. Яцко В.А. Компьютерная лингвистика или лингвистическая информатика? // Научно-техническая информация. Сер. 2. – 2014. – № 5. – С. 1–10.
2. Köhler R., Rieger B.B. Preface // Contributions to quantitative linguistics. Proceedings of the First international conference on quantitative linguistics. – Dordrecht, 1993. – P. i–xi.
3. Михайлов А.И., Черный А.И., Гиляревский Р.С. Информатика – новое название теории научной информации // Научно-техническая информация. – 1966. – № 12. – С. 35–39.
4. Piantadosi S.T. Zipf's word frequency law in natural language: a critical review and future directions. – 2014. – URL: http://colala.bcs.rochester.edu/papers/ piantadosi2014zipfs.pdf.
5. Manning C.D., Raghavan P., Schütze H. An introduction to information retrieval. Online edition. – Cambridge (UK), 2009. – URL: http://nlp.stanford.edu/IR-book/ pdf/irbookonlinereading.pdf.
6. Altmann G., Popescu I.-I., Zotta D. Stratification in texts // Glottometrics. – 2013. – Issue 25. – P. 85–93.
7. Popescu I.-I., Mačutek J., Altmann G. Aspects of Word Frequencies. Lüdenscheid: RAM-Verlag, 2009. – 193 p.
8. Gabaix X. Zipf's law for cities: an explanation // The quarterly journal of economics. – 1999. – Vol. 114, № 3. – P. 79–767.
9. Novovičová J., Malik, A. Information-theoretic feature selection algorithms for text classification // Proceedings of international joint conference on neural networks. – Montreal, 2005. – P. 3272–3277. – URL: http://staff.utia.cas.cz/novovic/files/1483.pdf.
10. Nicolosi N. Feature selection methods for text classification. – URL: http://www.cs.rit.edu/~nan2563/feature_selection.pdf.
11. Oakes M. P., Gaizauskas R., Fowkes H. A method based on the chi-square test for document classification // SIGIR '01 Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval. – New York, 2001. – URL: http://pers-www.wlv.ac.uk/~in4326/ old/2001_Oakes_SIGIR.pdf.
12. Яцко В.А. Метод зонального анализа данных // В мире научных открытий. – 2013. – № 6.1. – С. 166–182.
13. Яцко В.А. Метод зонально-корреляционного анализа текста // Научно-техническая информация. Сер. 2. – 2014. – № 10. – С. 26–30; Yatsko V.A. The Method of Zonal Correlation Text Analysis // Automatic Documentation and Mathematical Linguistics. – 2014. – Vol. 48, № 5. – P. 259–263.
14. West M. The mystery of Zipf. – URL: http://plus.maths.org/content/mystery-zipf.
15. Ahlgren O., Malo P., Sinha A, et al. A dimensionality reduction approach for semantic document classification. – URL: http://iswc2011.semanticweb.org/fileadmin/iswc/Papers/Workshops/SPIM/spim2011_paper6.pdf.

Материал поступил в редакцию 17.02.15.

Сведения об авторе

ЯЦКО Вячеслав Александрович – доктор филологических наук, профессор, Хакасский государственный университет им. Н.Ф. Катанова, г. Абакан
e-mail: viatcheslav-yatsko@rambler.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’32 : [001.4 : 62]

Н.А. Кочеткова

Метод извлечения технических терминов с использованием усовершенствованной меры странности

Представлен метод извлечения терминов из текстов технической области, основанный на новой мере терминологичности. Для отбора кандидатов в термины используются морфологические ограничения. Приводятся результаты экспериментов на корпусе текстов по тематике «Системы автоматизации проектирования и компьютерная графика».

Ключевые слова: *извлечение терминов, меры терминологичности, извлечение знаний из текстов*

ВВЕДЕНИЕ

В современном мире наблюдается постоянный рост объемов информации и активное развитие новых направлений и областей знания. Создание систем автоматизированного анализа данных является одной из самых актуальных задач в сфере информационных систем. Для корректной работы современные системы анализа данных должны содержать актуальные, обновляемые словари терминов каждой предметной области, в которой будут использоваться. Терминологические словари необходимы для решения многих задач анализа текстов на естественном языке. К таким задачам относятся индексирование, автореферирование, классификация документов, извлечение знаний и информационный поиск. Также при проектировании систем анализа данных требуется составление справочника терминов, который должен использоваться для точного определения и фиксации списка используемой заказчиком, проектировщиком и пользователем системы понятий, а также для создания тезауруса, выделения списка проектируемых или моделируемых сущностей, входящих в состав системы, и других подобных задач.

Очевидным решением здесь является создание словарей терминов предметной области на основе уже существующих терминологических словарей. К сожалению, этот подход невозможен для целого ряда областей. Так, существуют обширные общие словари инженерных терминов [1] и очень небольшие словари по более узким направлениям, например, по программированию [2] или искусственному интеллекту [3]. Словарь [3] содержит всего 550 терминов, которые не способны полностью описать область искусственного интеллекта. Кроме того, подобные словари стремительно теряют свою актуальность с момента издания из-за возникновения новых понятий. Для новых областей знания словарей терминов на русском языке в подавляющем большинстве случаев не существует вовсе, так как выделение терминов из текста вручную – крайне трудоемкий и долгий процесс. В связи с этим возникает задача автоматизации процессов выделения терминов из корпусов текстов для составления терминологических словарей.

На данный момент существует множество работ, посвященных автоматизированному извлечению терминов для различных языков. Так в работах [4–6] приведен сравнительный анализ для английского, итальянского и французского языков, соответственно. Методы, наиболее эффективные для русского языка, и их подробная оценка приведены в работах [7] и [8]. В работе [7] показано, что даже простая частота встречаемости может давать показатели не хуже, чем такие меры как LR и χ^2. В работе [8] среди статистических мер терминологичности лучшие результаты продемонстрировали c-value и weirdness. Также в работе [7] отмечено, что особую сложность для оценки представляют технические тексты при сравнительной простоте научных текстов в целом: показатели согласия экспертов составили 44% – для книги «Сетевые операционные системы» и 77% – для книги «Философия. Методология. Наука». Это означает, что постобработка словарей, полученных для технической области, является более сложной. Кроме того, значение F-меры для результатов работы c-value и weirdness на корпусе, например, новостных текстов колеблется в пределах 50–60%, что также означает, что постобработка полученных таким образом словарей будет трудоемкой.

Стоит отметить, что в качестве тестовых корпусов в работе [8] были выбраны новостные статьи и гуманитарные журналы. В связи с этим вновь возникает вопрос эффективности применения описанных в ней методов к техническим текстам.

Исходя из изложенного, целью данного исследования является повышение эффективности выделения терминов из текстов технической направленности за счет повышения точности и полноты выделенных терминов.

СУЩЕСТВУЮЩИЕ РЕШЕНИЯ

Процесс автоматизированного получения терминологического словаря из текстов предметной области обычно состоит из трех этапов:

1) формирование текстового корпуса;
2) применение методов извлечения терминов;
3) проверка и исправление результатов экспертами.

Задача формирования репрезентативных корпусов предметных областей в данной статье рассматриваться не будет, например, потому, что зачастую при анализе текстов технической области приходится работать лишь с их ограниченным набором, а источники новых текстов могут отсутствовать.

Большинство существующих методов извлечения терминов из текста работают в два шага. На первом шаге извлекаются все возможные кандидаты в термины. На втором – осуществляется фильтрация истинных терминов. Более подробно они будут рассмотрены при описании предлагаемого метода.

Недостатком многих существующих методов является тот факт, что они посвящены выделению терминов заданной, причем небольшой, длины: однословных [5] или двухсловных [9]. Так же часто встречается подход, отбирающий максимально длинные термины [6]. На практике же термины не имеют ограничения по длине (например, «система управления хладогенератора второго круга системы пуска установки»).

Вслед за [7] нас интересует метод, который может быть использован для выделения терминов произвольной длины. В связи с этим широко распространённые статистические меры в данном случае не применимы (не существует формул MI, t-score для терминов, состоящих более чем из трёх слов [10]). Более того, методы, отдающие предпочтение более частотным кандидатам, такие как t-score, не работают при увеличении длины термина, так как при увеличении длины частота встречаемости термина падает.

Кроме того, для выделения терминов произвольной длины требуется не накладывать строгих ограничений на их структуру. Это означает, что на этапе выделения терминов-кандидатов необходимо использовать минимум информации о структуре и составе терминов.

МЕТОД ВЫДЕЛЕНИЯ ТЕРМИНОВ

Основу предлагаемого метода составляет утверждение, что большинство терминов представляют собой именные группы, которые часто встречаются в текстах заданной предметной области и редки или отсутствуют в текстах других областей (более подробное исследование см., например, в [11]).

Этап отбора

На этом этапе из текста выделяются все возможные конструкции, которые могут представлять собой термины. К самым распространенным методам выделения кандидатов в термины относятся лексико-синтаксические шаблоны, морфологические шаблоны и коллокации.

Лексико-синтаксический шаблон [12] – шаблон, задающий характерные конструкции терминов. Шаблоны определяют лексемы конструкции и их грамматические параметры, а также задают синтаксические условия заполнения пропусков в конструкции. Для реализации данного подхода необходимы полные лингвистические знания синтаксических конструкций и ключевых слов извлекаемых терминов.

Коллокация – это комбинация двух или более слов, имеющих тенденцию к совместной встречаемости [10]. Статистическая модель коллокаций соответствует скрытой Марковской цепи порядка n-1. Появление больших репрезентативных корпусов текстов позволяет получить достоверные данные о частоте сочетания в языке в целом. Этот метод крайне прост в реализации, но требует тщательного составления исходного корпуса и последующей статистической обработки. Использование этого подхода для извлечения кандидатов в термины проигрышно, так как заведомо отсеиваются все редкие (t-score) или самые частотные (MI) кандидаты [10].

Метод морфологических шаблонов [7] основывается на предположении, что большинство кандидатов в термины являются фиксированными последовательностями морфологических единиц, т.е. в основе лежит та же идея, что и в лексико-синтаксических шаблонах, но рассматриваются только части речи и грамматические параметры, что позволяет избежать создания специального языка (лексемы и типы синтаксических связей не учитываются). Существующие реализации этого метода для выделения кандидатов в термины основываются на предположении о том, что большинство терминов являются именными группами. Это предположение подтверждается, например, в статье [11].

Для реализации метода морфологических шаблонов необходим морфологический анализатор и шаблоны. Вместо задания списка правил, содержащих описание грамматических характеристик слов, составляющих термин, в настоящей работе было решено наложить ограничения лишь на части речи. В нашем случае термины могут состоять только из существительных, прилагательных, причастий, порядковых числительных, предлогов и союза «и». Наречия и местоимения не разрывают термин, но и не входят в его состав, т.е. игнорируются. Такое ограничение позволяет нам получать термины произвольной длины, без предварительных исследований характерных конструкций. Слова, составляющие извлекаемые словосочетания, приводятся к нормальной форме.

Этап фильтрации

На первом шаге фильтрации отсеивались все сочетания, частота встречаемости которых в корпусе была меньше трёх. Хотя некоторые термины в корпусе могли встретиться всего один-два раза, и, следовательно, не пройти этот фильтр, это, скорее, вопрос формирования репрезентативного корпуса.

В настоящей работе существующие терминологические словари не были использованы из-за их недостаточного объема. По этой же причине они не использовались для определения полноты полученного списка требований или при машинном обучении. Также остались без внимания лексические методы извлечения терминов, так как они требуют предварительного исследования текстов предметной области и написания для нее шаблонов, задающих термины.

Среди статистических методов самыми перспективными по оценкам, приведенным в работах [4, 7], являются C-value, tf-idf и weirdness.

C-value – метод выделения многословных терминов[13]. Так как встречаемость длинных терминов в тексте меньше, чем коротких, то для компенсации этого эффекта метод C-value поощряет словосочетания, не входящие в состав других, более длинных. Значение меры C-value рассчитывается следующим образом:

$$\text{C-value}(a) = \begin{cases} \log_2|a| * \text{freq}(a), \text{ если не вложен} \\ \log_2|a| * \text{freq}(a) - 1/N(Ta) * \sum \text{freq}(b), \end{cases}$$

где: a – кандидат в термины,

|a| – длина словосочетания, измеряемая в количестве слов,

freq(a) – частотность a,

Ta – множество словосочетаний, которые содержат a,

N(Ta) – количество словосочетаний, содержащих a,

$\sum$freq(b) – сумма частот всех сочетаний, содержащих a,

b – сочетания, содержащие в себе a.

Чем больше частота термина-кандидата и его длина, тем больше его вес. Если этот кандидат входит в большое количество других словосочетаний, его вес уменьшается.

Однако в технических текстах части термина зачастую сами по себе являются терминами. Для примера приведем частоты и C-value для терминов, содержащих сочетание «программное обеспечение». Термин «программное обеспечение» имеет частоту 272 и c-value, равную 175, термин «разработчик программного обеспечения» – 12 и 19 соответственно, а «лицензионное программное обеспечение» – частоту 6 и c-value 10. Хотя все они являются терминами, последний получает слишком низкую оценку из-за недостаточно высокой частоты.

Мера Tf*idf позволяет снижать вес общеупотребимых слов. Tf – это частота слова в корпусе. Idf – обратная подокументная частота слова.

$$Tf = n_i/N,$$

где n_i есть число вхождений слова в документ, а N в знаменателе – общее число слов в данном документе.

$$idf = |d_i \supset t_i| / D,$$

здесь D – количество документов в корпусе, $|d_i \supset t_i|$ – количество документов, в которых встречается t_i (когда $n_i \neq 0$).

$$Tf*idf(t, d, D) = tf(t, d) \times idf(t, D).$$

Странность (Weirdness) [14] – мера, учитывающая пропорциональное соотношение частотности употребления слова в рабочей текстовой коллекции по сравнению с контрастной коллекцией.

Пусть w – слово.

Тогда

$$\text{Weirdness}(w) = (Ws / Ts) / (Wg / Tg),$$

где: Ws – частотность слова в коллекции предметной области;

Ts – общее количество словоупотреблений в коллекции предметной области;

g – контрастная коллекция;

Wg – частотность слова в контрастной коллекции;

Tg – общее количество словоупотреблений в контрастной коллекции.

В классическом варианте возможно Wg = 0; чтобы избежать этого мы модифицировали формулу:

$$\text{Weirdness}(w, g) = (Ws / Ts) / ((Wg + Ws) / (Tg + Ts)).$$

Вычисление мульти- и суперстранности

В основе предложенной идеи лежит предположение, что термином является частотное для предметной области сочетание, являющееся редким в текстах других областей (за исключением предметных областей, где данное слово или словосочетание также является термином, но несет другой смысл). То есть сочетание, имеющее высокую частоту и получающее высокое значение weirdness на всех корпусах, заведомо не относящихся к выбранной предметной области, точно является термином этой предметной области.

Данное утверждение можно усилить. Если использовать корпус аналогичного стиля, список терминов которого гарантированно имеет малое пересечение с заданным корпусом, то фильтрация относительно него позволяет отсеять стилистические маркеры, так как они получат невысокое значение странности.

Было решено проверить, в каком случае странность будет давать лучший результат: при использовании объединения двух контрастных коллекций или при их последовательном применении, т.е. рассматривалось два варианта: странность, подсчитанная относительно слияния двух контрастных коллекций, и при раздельном учете результатов по каждой контрастной коллекции с последующим перемножением результатов. Далее меру, равную произведению странностей, полученных на нескольких различных контрастных коллекциях, будем называть *мультистранностью.*

В дальнейших вычислениях использовалась формула странности относительно слияния двух коллекций 1 и 2:

$$\text{Weirdness}(w) = (Ws / Ts) / ((Wg1 + Wg2 + Ws) / (Tg1 + Tg2 + Ts)),$$

и формула для подсчета мультистранности относительно коллекций 1 и 2:

$$\text{MultiWeirdness}(w) = (Ws / Ts) / ((Wg1 + Ws) / (Tg1 + Ts)) * (Ws / Ts) / ((Wg2 + Ws) / (Tg2 + Ts)) = \text{Weirdness}(w,g1) * \text{Weirdness}(w,g2).$$

Здесь Ws – частотность слова в коллекции предметной области;

Ts – общее количество словоупотреблений в коллекции предметной области;

g1 – первая контрастная коллекция;

g2 – вторая контрастная коллекция;

Wg1 – частотность сочетания в первой контрастной коллекции;

Tg1 – общее количество словоупотреблений в первой контрастной коллекции;

Wg2 – частотность сочетания во второй контрастной коллекции;

Tg2 – общее количество словоупотреблений во второй контрастной коллекции.

Стоит отметить, что количество контрастных корпусов, необходимых для подсчета мультистранности, определятся лишь строгостью критериев: чем чище должен быть результат, тем больше потребуется убирать пересечений лексики, т.е. тем больше необходимо контрастных коллекций.

В общем случае мультистранность рассчитывается как произведение странностей на всех контрастных коллекциях:

$$MultiWeirdness(w) = \Pi(Weirdness(w, gi)) .$$

Так как многие термины принадлежат не одной, а сразу нескольким областям, иногда отличаясь по смыслу, эти термины получат невысокое значение как странности, так и мультистранности. Однако зачастую эти термины имеют длину, равную двум или одному слову, и довольно высокую частоту встречаемости в корпусе. Для того чтобы повысить их шансы на попадание в список терминов, было решено умножать значение мультистранности на частоту. Это позволяет сглаживать различия между общеупотребимыми и специфичными терминами, так как предполагается, что даже специфичные термины имеют высокую частоту в корректно составленном корпусе предметной области. Меру, равную произведению мультистранности и частоты встречаемости назовем суперстранностью.

$$SuperWeirdness(w) = MultiWeirdness(w)*freq(w),$$

где freq(w) — частота встречаемости сочетания в специальном корпусе.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Используемые данные

В качестве исходного корпуса были выбраны статьи, опубликованные в журнале «САПР и графика» с 2000 по 2013 гг. [15]. Корпус содержит примерно 4500 статей разных авторов, посвященных описанию различных вопросов автоматизированного проектирования систем, а также смежным областям, – общим объемом около 5 млн токенов. Корпус обрабатывался как единый текстовый документ, без учета разбиения на отдельные статьи. Морфологическая обработка осуществлялась с помощью программы Кросслейтор [16].

Следуя вышеописанным морфологическим ограничениям, из корпуса было извлечено 84915 уникальных групп существительного (кандидатов в термины). Затем сочетания с частотой встречаемости меньше трёх были отсеяны. В качестве контрастных коллекций были выбраны два корпуса: Библиотека Мошкова – около 560,3 млн. токенов и выборка журналов и книг по географической тематике (корпус «ГЕО») – около 3 млн токенов.

Выбор корпуса статей в качестве второй контрастной коллекции позволяет отфильтровать устойчивые публицистические обороты, присущие журналу «САПР и графика».

В ходе эксперимента были подсчитаны такие меры как c-value, tf-idf, weirdness относительно библиотеки Мошкова, weirdness относительно библиотеки Мошкова, объединенной с корпусом «ГЕО», мультистранность относительно библиотеки Мошкова и корпуса «ГЕО», суперстранность относительно библиотеки Мошкова и корпуса «ГЕО».

Эксперименты

Для сравнения методов проводилась оценка полученных списков терминов лингвистом и специалистом предметной области, а также проверка мер относительно друг друга.

Для оценки результатов лингвистом были выбраны 1000 сочетаний, получивших наибольшие значения суперстранности относительно выборки Librusec.ru за 2013 г. (около 840 млн. слов) и библиотеки Мошкова. Принимая во внимание только область САПР, было отобрано 415 терминов. Терминов САПР и смежных областей было отобрано 650. Таким образом, точность суперстранности на неподобранных, но больших, контрастных коллекциях составила 65%.

Эксперт вычитывал сочетания с наибольшими значениями меры суперстранности относительно Librusec.ru и библиотеки Мошкова до тех пор, пока не получал 1000 терминов. Было просмотрено 1608 пунктов. Точность составила 62%. Относительно этих отобранных списков терминов был подсчитан процент совпадений топ-1000 значений мер терминологичности. Результаты приведены в табл. 1.

Стоит отметить, что при расчете необходимо учитывать не только термины САПР, но и все встреченные термины смежных областей, так как именно такие смешанные словари и требуются для межпредметных областей знаний.

Как видно из таблицы, суперстранность показала лучшие значения, немного превосходя значения c-value – притом, что обе оценки имеют схожие показатели.

Имея семь мер терминологичности можно выявить их корреляцию. Для этого составим списки терминов, которые встретились как минимум в трех, четырех и пяти списках из списков топ-1000 значений мер. Данное сравнение показывает различие в извлекаемых сочетаниях. Стоит обратить внимание на то, что топ-1000 странностей, хоть и имеют одинаковую оценку точности, несколько различаются по своему содержанию. Количество совпадений приведено в табл. 2.

В табл. 3 показана доля терминов из списков совпадений в списках топ-1000 мер терминологичности.

Таблица 1

Процент совпадений мер терминологичности и отобранного топа суперстранности по Librusec.ru и библиотеке Мошкова

	Список, отобранный лингвистом	Список, отобранный экспертом
Частота	0,28	0,31
c-value	0,51	0,34
tf-idf	0,22	0,22
Weirdness относительно библиотеки Мошкова	0,18	0,13
Weirdness относительно объединения библиотеки Мошкова и журнала «ГЕО»	0,18	0,13
Мультистранность относительно библиотеки Мошкова и журнала «ГЕО»	0,18	0,14
Суперстранность относительно библиотеки Мошкова и журнала «ГЕО»	0,52	0,51

Таблица 2

Количество терминов, присутствующих хотя бы в трех, четырех или пяти и более списках топ-1000 мер терминологичности

	В трех и более списках	В четырех и более списках	В пяти и более списках
Количество совпавших	577	190	10

Таблица 3

Доля терминов, присутствующих хотя бы в трех, четырех или пяти и более списках топ-1000 мер терминологичности

	В трех и более списках	В четырех и более списках	В пяти и более списках
Частота	0,22	0,31	0,3
c-value	0,58	0,85	1
tf-idf	0,2	0,26	1
Weirdness относительно библиотеки Мошкова	0,97	1	1
Weirdness относительно библиотеки Мошкова и корпуса «ГЕО»	0,96	1	1
Мультистранность относительно библиотеки Мошкова и корпуса «ГЕО»	0,97	1	1
Суперстранность относительно библиотеки Мошкова и корпуса «ГЕО»	0,98	1	1

Данное сравнение наглядно демонстрирует отличие механизмов частотности и tf-idf от мер c-value и weirdness, а также высокий процент схожести у мер Weirdness, мультистранность и суперстранность – независимо от контрастных коллекций.

Для экспертной оценки метода суперстранности было отобрано 1000 сочетаний, получивших наибольшие значения суперстранности относительно библиотеки Мошкова и корпуса «ГЕО».

Экспертная оценка точности, в зависимости от строгости критериев терминологичности, колеблется от 52% (если считать термины, относящиеся только к САПР и графике и не относящиеся к общенаучным) до 81% (если считать все термины, которые употребляются в контексте САПР, включая экономические). Приведем список двадцати терминов, получивших максимальное значение суперстранности:

ПОЛЬЗОВАТЕЛЬ, СИСТЕМА, ПРОЕКТИРОВАНИЕ, ПРОЕКТИРОВЩИК, ПРОГРАММА, ИЗДЕЛИЕ, ПРЕДПРИЯТИЕ, МОДУЛЬ, ТЕХНОЛОГ, РАЗРАБОТЧИК, МОДЕЛЬ, КОНСТРУКТОР, САПР, ЧЕРТЕЖ, ЗАКАЗЧИК, ПРОЕКТ, ОБРАБОТКА, ПРОГРАММНЫЙ ПРОДУКТ, СБОРКА, ТЕХНОЛОГИЯ.

Для сравнения со словарем [3] было выбрано 50 самых частотных терминов по мультистранности относительно библиотеки Мошкова и корпуса «ГЕО». Только три из выбранных терминов встретились в словаре: «алгоритм», «интерфейс» и «модель». То, что термины, относящиеся к категориям «графика» и «САПР», отсутствовали в словаре по искусственному интеллекту, закономерно. Однако стоит отметить, что отсутствуют в нем и термины «компьютер» и «персональный компьютер», но при этом присутствует семейство статей вида «модель + *термин*».

Для оценки полноты экспертом были отчитаны 7 списков топ-1000 значений мер. Из всех терминов отчитанных списков был выделен общий список терминов, признанный «золотым стандартом». В нем оказалось 1315 терминов. Относительно него были рассчитаны полнота, точность и f-мера. Результаты приведены в табл. 4.

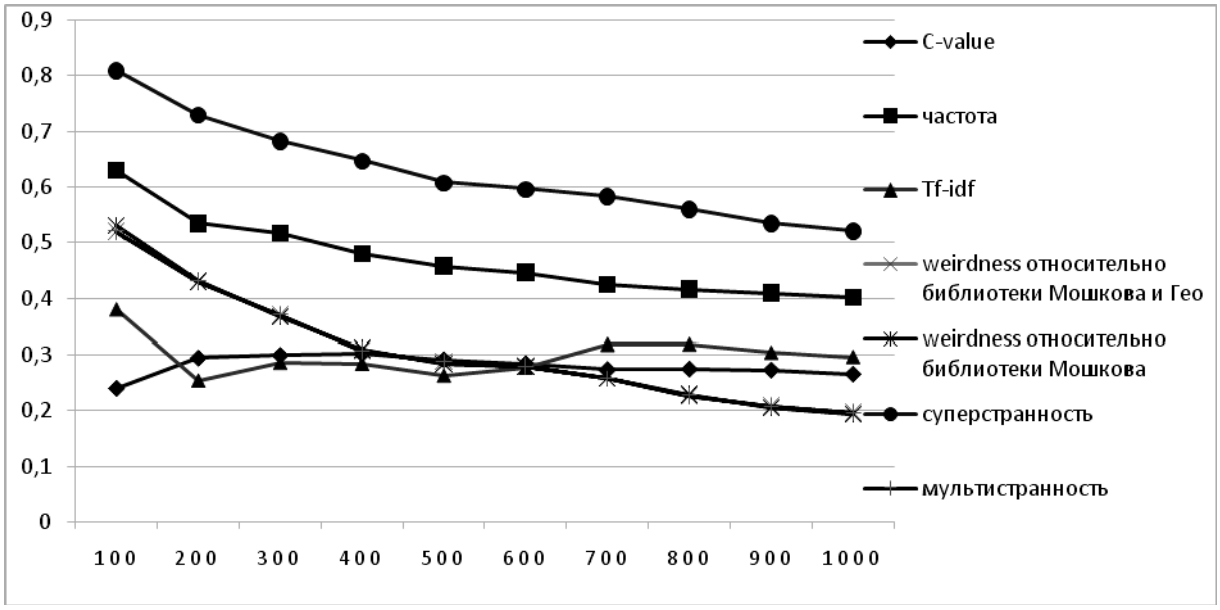
Изменение точности с увеличением количества рассматриваемых терминов на первой тысяче значений показано на рисунке.

Одинаковое поведение странностей и мультистранности на первой тысяче может быть объяснено тем, что наибольшие значения получили термины, не встретившиеся или встретившиеся один раз в контрастных коллекциях, и, следовательно, получившие практически одинаковые значения исходя из формулы. Так как суперстранность представляет собой произведение мультистранности на частоту встречаемости, а частота имеет тот же характер убывания, что и мультистранность, то на графике суперстранность имеет такой же вид, как и прочие странности.

Таблица 4

Показатели полноты и точности рассматриваемых методов

	Точность, %	Полнота, %	F-мера, %
Частота	40	30	34
c-value	23	17	20
tf-idf	34	20	25
Weirdness относительно библиотеки Мошкова	19	16	17
Weirdness относительно библиотеки Мошкова и корпуса «ГЕО»	19	16	17
Мультистранность относительно библиотеки Мошкова и корпуса «ГЕО»	20	18	19
Суперстранность относительно библиотеки Мошкова и корпуса «ГЕО»	54	41	47



Графики изменения точности при увеличении количества рассматриваемых пунктов

Отличный вид tf-idf объясняется предпочтением более редких терминов. Поэтому в первую сотню tf-idf попали термины, встреченные много раз, но в одном тексте, такие как «ландшафтный дизайн», «растение», «нефтепровод». В первой сотне c-value преобладают общеупотребимые высокочастотные сочетания: «в первую очередь», «новая версия», «результат расчета», «в качестве примера», «особое внимание».

ЗАКЛЮЧЕНИЕ

Анализ результатов экспериментов показывает более высокую, по сравнению с существующими методами, точность предложенного метода выделения терминологии для новых областей знания. Основной особенностью метода суперстранности является способность поэтапно отсеивать ненужную лексику. То есть при использовании этого метода для выделения терминов из технической документации проводится отсеивание общей лексики по библиотеке Мошкова, а лексики, связанной с характерными для документации оборотами, – по корпусу технической документации другого направления. Суперстранность включает в себя два вида терминов – как странные, так и наиболее частотные, что позволяет получать бóльшую полноту по сравнению с другими методами.

Использование меры суперстранности позволяет выделять большинство терминов заданного корпуса. Она позволяет получать междисциплинарные словари терминов, а также словари терминов для новых областей. Точность метода на выбранном корпусе составила не менее 52%.

Часть ошибок в проведенных экспериментах объясняется не совсем чистым корпусом предметной области: в журнале имеются статьи экономического характера, а также презентации новых программных продуктов. Если не исключать привносимую ими терминологию, то точность может быть оценена в 81%.

Повысить точность метода можно применением списков стоп-слов. К самым распространенным стоп-словам для технических текстов относятся: «выбор», «большинство», «возможность», «информация о», «изменение». Также при повышении порога частоты до пяти наблюдается существенное улучшение качества (порядка 5%).

Особую проблему составляют «термины-прилипалы», такие как «качество», «тип», «модель», после которых почти всегда следует уточняющий термин (напр., «тип фильтра»). Однако эти термины встречаются во многих областях и могут быть занесены в списки стоп-слов (или термины, содержащие их, – выделены в отдельный список для более тщательной проверки экспертом) или могут быть отнесены к терминам целиком.

В ходе дальнейших работ планируется произвести оценку метода для текстов других предметных областей, а также для текстов на других языках. Однако для оценки результатов потребуется эксперт в выбранной предметной области.

Итак, предложенный метод пригоден для использования в автоматических системах извлечения информации, однако, для составления словарей требуется вычитка полученных терминов экспертом. По сравнению с другими методами суперстранность позволяет регулировать содержание словаря, а высокая точность существенно сокращает время работы экспертов.

СПИСОК ЛИТЕРАТУРЫ

1. Словарь инженерных терминов. – URL: http://inzhenery.su/slovar/.
2. Информатика. Словарь основных терминов. – URL: http://book.kbsu.ru/theory/definition.html.
3. Толковый словарь по искусственному интеллекту. – URL: http://www.raai.org/library/tolk/aivoc.html.
4. Lars Ahrenberg. Term Extraction: A Review, 2009. – URL: http://www.ida.liu.se/~lah/ Publications/ tereview_v2.pdf.
5. Francesca Bonin, Felice Dell'Orletta, Giulia Venturi, Simonetta Montemagni. A Contrastive Approach to Multi–word Term Extraction from Domain Corpora // Proceedings of the LREC 2010, Seventh International Conference on Language Resources and Evaluation, 2010. – P. 3222–3229.
6. Bourigault D. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases // Proceedings of the COLING-92, 1992. – P. 977–981.
7. Браславский П.И., Соколов Е.А. Сравнение пяти методов извлечения терминов произвольной длины // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. Диалог'2008. – М.: Изд. РГГУ, 2008. – С. 67–74.
8. Лукашевич Н.В., Логачев Ю.М. Использование методов машинного обучения для извлечения слов-терминов // Одиннадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2010. – С. URL: http://www.raai.org/resurs/papers/kii-2010/ doklad/loukachevitch.doc.
9. Браславский П., Соколов Е. Сравнение четырех методов автоматического извлечения двухсловных терминов из текста // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. Диалог'2006. – М.: Изд-во РГГУ, 2006. – С. 88–94.
10. Хохлова М.В. Экспериментальная проверка методов выделения коллокаций // Сб. статей «Инструментарий русистики: корпусные подходы». – Хельсинки, 2008. – С. 343–357.
11. Шелов С.Д. Семь вопросов и семь ответов по семантике термина. – М.: ВИНИТИ, 2001. – С. 72–83.
12. Большакова Е.И., Баева Н.В., Бордаченкова Е.А., Васильева Н.Э. Морозов С.С. Лексико-синтаксические шаблоны в задачах автоматической обработки текста // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции Диалог'2007. – М.: Изд-во РГГУ, 2007. – С. 70–75.

13. Ananiadou S. A methodology for automatic term recognition // Proceedings of COLING-1994. – P. 1034–1038.

14. Ahmad K., Gillam L., Tostevin L. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder) // Eighth Text Retrieval Conference (TREC-8), 1999.

15. САПР и графика. – URL: //http://www.sapr.ru/.

16. Клышинский Э.С., Елкин С.В., Стеклянников С.Е. Проблемы создания универсального морфосемантического словаря // Сб. трудов Международных конференций IEEE AIS'03 и CAD-2003. Т. 1. – Дивноморское, 2003. С. 159–163.

Материал поступил в редакцию 13.02.15.

Сведения об авторе

КОЧЕТКОВА Наталия Александровна – аспирант Московского института электроники и математики Национального исследовательского университета «Высшая школа экономики», Москва
e-mail: natalia_k_11@mail.ru

СПРАВОЧНО-ИНФОРМАЦИОННЫЙ РАЗДЕЛ

УДК 001.102 : 004

Е. А. Плешкевич

На пути к информационной картине мира Джеймса Глика*

ВВЕДЕНИЕ

Одной из актуальных задач современной науки является построение модели информационной картины мира, как особой формы знания о нем, основанного на представлениях о феномене информации. В решении этой задачи задействованы не только ученые, но и писатели, в чью задачу входит популяризация научных исследований. Одним из шагов в этом направлении стала публикация в 2011 г. монографии американского писателя Джеймса Глика «Информация. История. Теория. Поток» (далее – «Информация») и ее перевод в 2013 г. на русский язык. По версии *Los Angeles Times* она была признана лучшей научной книгой 2011 г. В следующем году «Информация» была отмечена рядом наград, среди которых мы хотели бы выделить премию Королевского научного общества, по мнению жюри которого – это одна из тех редких книг, которые дают совершенно новую основу для восприятия мира [1]. Ее автор, известный американский писатель, популяризатор науки Джеймс Глик (James Gleick) родился в 1954 г. В 1976 г. он окончил колледж в Гарварде по специальности «Английский язык и литература». С 1979 по 1989 гг. работал журналистом *New York Time.* Из-под его пера вышли книги, посвященные Исааку Ньютону (2003) и Ричарду Фейнману (1992), а также теории хаоса (1987). Нам представляется, что обращение к теме хаоса – физического явления, рассматриваемого в ряде случаев как антипода информации, во многом предопределили последующее обращение писателя к теме информации, вернее, к картине мира, в которой информация и хаос являются его центральными персонажами.

Прежде чем перейти к обзору «Информации», хотим подчеркнуть, что монография носит научно-популярный характер. Ее цель – познакомить с информационной проблематикой широкий круг читателей, включая как специалистов, так и тех, кто случайно взял эту книгу в руки. Это, во-первых. Во-вторых, книга информативна, содержит множество интересных фактов и авторских интерпретаций, которые мы, в силу ограниченности жанра рецензии-обзора, в надежде на то, что наша публикация стимулирует интерес читателей, как к феномену информации, так и к самой книге, вынуждены оставить за скобками. Таким образом, основную цель мы видим в оказании содействия всем тем, кто вместе Джеймсом Гликом собирается отправиться в увлекательное путешествие по миру информации и информационных технологий.

ОТ СЛОВА К ТЕЛЕГРАФУ

В качестве отправной точки своего виртуального путешествия Джеймс Глик предлагает выбрать Слово. В начале было Слово, и Слово было у Бога, и Слово был Бог – именно так начинается Евангелие от Иоанна. Именно письменное слово, а не огонь, отмечает Эсхил в своей трагедии «Прометей прикованный», на самом деле было величайшим даром Прометея человечеству: «Премудрость чисел, из наук главнейшую, я для людей измыслил и сложенья букв, мать всех искусств, основу всякой памяти». Рассуждая о значении письменного слова, автор обращается к исследованиям американского философа и историка культуры Уолтера Дж. Онга (1912–2003 гг.). В них ученый подчеркивает, что значение письменности заключается не только в существенном расширении коммуникационных возможностей человека, когда, по выражению Платона, один может говорить с множеством, мертвый – с живыми, а живые – с еще не родившимся. Оно существенно возрастает, если принять во внимание изобретение греческой культурой формальной логики как способа мышления, что стало возможным только благодаря использованию алфавитного письма. При этом телефон, факс, калькулятор и, в конечном счете, компьютер – всего лишь изобретения, которые нужны для хранения, использования и передачи информации и знаний, созданных на основе письменного слова. Безусловно, что история становления информационных технологий началась не с телефона и телеграфа. Она началась с упорядочения информационного пространства, по крайней мере, той его части, которая была образована письменно-

* Рец. на кн.: Gleik J. The Information. A History. A Theory. A Flood. – NY : Pantheon Books, 2011. – 526 p.; Глик Дж. Информация. История. Теория. Поток / пер. с англ. М. Кононенко. – М. :АСТ : CORPUS, 2013. – 576 с.

стью. Одним из первых шагов в этом направлении стало составление лексических словарей и сортирования в них слов в алфавитном или ином порядке. Обращаясь к их истории, Глик отмечает, что, возможно, первые лексические словари впервые появились еще в Александрийской библиотеке. Мы полагаем, что они имели место и ранее, скажем, в Древней Месопотамии, однако, если принимать во внимание именно фонетическую письменность, то в этом случае отсчет следует вести, конечно, с Александрийской библиотеки. На этом пути автором выделяются изобретение монахом-доминиканцем Иоганном Бальбом в 1286 г. алфавитного порядка расположения слов, составление Робертом Кодри в 1604 г. "A Table Alphabetical…" т.е. «Перечня алфавитного …». Для сравнения отметим, что первый русский печатный словарь «*Лексис, сиречь речения вкратце собранны и из словенского языка [старо-церковного] на просты русский диялект истолкованы*» Лаврентия Зизания Тустановского объемом в 1061 слово был опубликован в Вильно (Вильнюсе) в 1596 г. В 1627 г. в Киеве вышел гораздо больший по объему (около 7 тыс. слов) «*Лексикон славено-русский и имен толкование*» Памвы Берынды, переизданный в 1653 г. и оказавший значительное влияние на последующие отечественные словари. В качестве еще одного шага автор выделяет разработку смыслового (предметного) принципа лексического упорядочения Г. В. Лейбница, связанного с движением мысли от предмета к слову, а не наоборот, как в случае алфавитного принципа упорядочения. Квинтэссенцией словарного развития английского языка, по мнению автора, стал Оксфордский словарь английского языка, задуманный Лондонским филологическим обществом в 1857 г., включающий сегодня все слова, бытующие или бытовавшие в английском литературном и разговорном языке, начиная с 1150 года.

Следующая информационная технология, как отмечает Глик, была связана с вычислениями, с повышением их эффективности. Одной из первых технологий на этом пути стало изобретение математических цифровых таблиц, содержащих ранее произведенные вычисления, как простых арифметических действий (умножение, деление), так и более сложных математических функций (возведение в степень или нахождения корней). Одной из самых известных математических таблиц, стала таблица умножения Пифагора (ок. 570–500 г. до н. э.). Впоследствии число таких таблиц постоянно расширялось. Например, в 1582 г. Симон Стевин составил сборник таблиц расчета процентов для банкиров и ростовщиков. Позже к ним добавились логарифмические таблицы, используемые при расчетах в астрономии и мореплавании. Однако таблицы, вычисляемые вручную, содержали множество ошибок, которые негативно влияли, например, на безопасность мореплавания. Выход из сложившейся ситуации виделся в постепенной механизации вычислений. Первыми такими механическими устройствами был абак – счетная доска, сконструированная (ориентировочно) еще в V веке до н.э. Впоследствии к нему добавились сконструированные в 1642 и 1672 гг. арифмометры Блеза Паскаля и Готфрида Лейбница. Однако своего пика идея механизации вычислений достигла в аналитической машине, предложенной английским математиком Чарльзом Бэббиджем (1791–1871 гг.). Именно в ней, пусть и виртуальной, по мнению Глика, информация была впервые представлена в виде чисел и процессов, которые должны были проходить через определенные физические точки, названные Бэббиджом складом для хранения (store) и мельницей для действия (mill). Сегодня мы бы сказали, что виртуальная аналитическая машина была наделена памятью, хранящей данные, и процессором, выполняющим расчеты. Первая программа для этой машины была составлена дочерью Байрона математиком Адой Лавлейс (1815–1852 гг.). Идея создания аналитической машины, несмотря на то, что она не была в тот момент реализована на практике, в целом была воспринята положительно. Известный американский писатель Эдгар Аллан По писал, что вычислительная машина г-на Бэббиджа, созданная из дерева и металла, может дать математически точные результаты операций, потому что способна исправлять собственные возможные ошибки.

Изобретение электричества, вернее, технологий по его использованию для передачи сообщений, по мнению Глика, кардинально повлияло на развитие информационных технологий и привело не только к созданию новых телекоммуникационных устройств, но и создало основу для появления первых теоретических представлений об информации. Напомним, что на основе электрического тока в 1844 г. заработал электрический телеграф, в 1876 г. – телефон, а в 1895 г. – радио. Новые провода – новая логика!

ИНФОРМАЦИОННАЯ ТЕОРИЯ ТЕЛЕГРАФНЫХ ПРОВОДОВ

Что же поменялось с изобретением телеграфа? Во-первых, обладание машинами для передачи и приема сигналов стало восприниматься как нечто естественное. Напомним, что еще совсем незадолго до этого Джонатан Свифт в «Путешествии Гулливера», высмеивал проект лапутянского профессора по усовершенствованию умозрительного знания при помощи технических и механических операций. Во-вторых, инженерные разработки в области электрической связи стимулировали разработку новых знаний о процессе передачи сообщений. Появилась необходимость введения нового теоретического понятия, позволяющего описывать процесс передачи текстового сообщения по каналам электрической связи. В качестве такового был выбран термин «информация» и предложено его новое трактование[1]. Оно было связанно с рассмотрением информации как величины, наполненной определенным физическим содержанием. И, наконец, в-третьих, была разработана технология сведения всего разнообразия алфавита к «точке», «тире» и паузе. Тем самым был сделан шаг к двоич-

[1] Напомним, что до этого термин информация трактовался как сведения, сообщения или данные.

ному коду, широко распространенному в современной компьютерной технике.

Как отмечает Глик, пионерами новой информационной логики в 1920-1940-х гг. стали американские инженеры и ученые – Гарри Найквист (1889–1976 гг.), Ральф Хартли (1888–1970 гг.), Клод Шеннон (1916–2001 гг.) и Уоррен Уивер (1894–1978 гг.). В начале 1920-х гг. Гарри Найквист предложил преобразовать аналоговый электрический сигнал в дискретный или цифровой сигнал, таким образом преобразовав его в конечное, счетное множество.

Следует отметить, что примерно над той же проблемой в СССР в 1930-х гг. работал В. А. Котельников (1908–2005 гг.). В конце 1920-х гг. Ральф Хартли в статье, посвященной передаче информации [2], предложил, во-первых, трактовать информацию как неопределенность или неожиданность, связанную с выбором символов из заранее заданного репертуара. Во-вторых, разработал один из первых алгоритмов расчета меры информации. В ее основу, как он сам заявляет, практической мерой информации был положен логарифм числа возможных последовательностей символов

$$H = n \bullet \log(S),$$

где H – количество информации, n – число переданных символов, S – размер алфавита.

Эта формула позволяет вычислить количество информации, содержащейся в сообщении, состоящем из n-числа символов, входящих в S-алфавит[2] . Логарифмический характер этих отношений был установлен Хартли эмпирически в ходе анализа печатающей телеграфной системы типа Бодо, где оператор выбирает буквы и другие знаки, каждый из которых при передаче состоит из последовательности обычно пяти символов. Для того чтобы мера информации, отмечает Хартли, имела практическую, инженерную ценность, она должна быть такой, чтобы информация была пропорциональна числу выборов.

Следующий шаг в формировании информационной логики электрической связи был сделан К. Шенноном и У. Уивером. В работе, посвященной математической теории связи [3], Шеннон развил логарифмический подход к определению количества информации, акцентировав внимание на двоичном логарифме, поскольку это было удобно для расчетов устройств, работающих на основе реле, имеющего, как известно, два состояния и ввел меру измерения информация, известную сегодня как «бит». Что касается нового толкования информации, то Шеннон также предложил рассматривать ее в интересах инженерной науки исключительно в физическом контексте. Что касается сущности информации, то, по мнению Шеннона, информация есть энтропия, само по себе трудное и плохо понимаемое понятие, обозначающее меру неупорядоченности систем в термодинамике, науке о температурах и энергии. Вернее, как отмечал в свое время А. Д. Урсул, комментируя идеи Шеннона, эта та часть энтропии, которая исчезает после получения сообщения [4, с. 62]. Он усовершенствовал формулу Хартли. Теперь мера информации могла быть определена по следующей формуле.

$$H(x) = -\sum_{i=1}^{n} p(i) \log_2 p(i) \ .$$

где p – функция вероятности[3].

Эта формула Шеннона, по наблюдениям Урсула, не уступает известности знаменитой формуле Эйнштейна, в которой энергия вещества оказалось равной произведению массы на скорость света в квадрате. Кроме этого Шеннон разработал модель передачи информации, включающую сигнал, передающий его передатчик, канал связи, на который воздействуют шумы, прием сигнала посредством приемника и передачу его получателю. В дальнейшем идеи Шеннона получила развитие в исследованиях У. Уивера, что привело к включению его имена в название теории «Шеннона–Уивера» [5].

Отказ от рассмотрения семантических аспектов информации, связанных с содержанием передаваемого сообщения, был подвергнут критике. Исключение «смысла» из понятия информация, заявляет Хайнц фон Ферстер (1911-2002 гг.), позволяет идентифицировать теорию информации Шеннона как теорию сигналов, говорящей о «бип-бипах», а не об информации (с. 266, 443)[4]. В результате критики чисто инженерного подхода к исследованию информации, отмечает Глик, в исследованиях наметился определенный поворот. В нем приняли участие американский математик и основатель кибернетики Норберт Винер (1894-1964 гг.) [6], Х. Ф. Ферстер и ряд других ученых. Обращаясь к Винеру, Глик отмечает, что тот, придерживается энтропийного подхода, однако предлагает иную трактовку энтропии, связывая ее не с неопределенностью как Шеннон, а с беспорядком. В новом контексте Ферстер начинает рассматривать информацию, как порядок, вычисленный из беспорядка (с. 266). В кибернетике Винер связывает информацию с саморегуляцией, с так называемым феноменом «обратной связи», при котором результаты, полученные на выходе системы, начинают учитываться системой. Заявляя о том, что информация есть информация, а не материя и не энергия, Винер утверждает, что рассмотрение информации в системах электрической связи как чисто материального про-

[2] Допустим,мы используем алфавит (m) состоящий из двух букв «Y» и «X», на основе которого составляем сообщение, состоящее (n) из трех букв. Таким образом, m=2, n=3. На этой основе можно составить восемь разных сообщений(N) «YYY», «YYX», «YXY», «YXX», «XYY», «XYX», «XXY», «XXX». Математически это можно представить следующим образом: $N = m^n = 2^3 = 8$.

[3] Так, количество информации, которое содержит слово «кот» если исходить из его написания буквами русского алфавита (32 буквы) может быть определено следующим образом. Каждая буква алфавита, исходя из двоичного логарифма, используемого в формуле Шеннона и репертуара русского алфавита состоящего из 32 букв, оказывает равной 5 битам, информационная емкость самого же слова «кот»,состоящего из 3 букв, оказывается равной 15 битам.

[4] Здесь и далее по тексту в круглых скобках указаны страницы рецензируемого издания на русском языке.

цесса, включающего физический процесс и энергетические затраты связанные с его реализацией, не раскрывают сути информации в живой природе и обществе. Под информацией он предлагает понимать обозначение содержания, полученное нами из внешнего мира в процессе приспособления к нему нас и наших чувств. Таким образом, как нам представляется, противостояние по линии «Шеннона – Винера» во многом обусловлено различиями в предмете их исследований. Если Шеннон и его сторонники разрабатывали понятие информации в целях определения необходимых и достаточных ресурсов для осуществления коммуникации по системам электрической связи, то объектом исследования Винера была информация как ресурс, обеспечивающий эволюцию и развитие в живой природе и социуме.

Одну из попыток математического определения информации предприняли в 1960-х гг. американский математик Рэй Соломонофф (1926–2009 гг.) [7] и советский академик А. Н. Колмогоров (1903–1987 гг.) [8, 9]. Они предложили алгоритмический подход, который, в отличие шенноновского подхода, был сосредоточен не на множестве, а непосредственно на объекте. Опираясь на понятие условной энтропии объекта, Колмогоров предложил ее рассматривать как минимальную длину записанной в виде последовательности нулей и единиц «программы», которая позволяет построить объект ***«x»***, имея в своем распоряжении объект ***«y»*** [8, с. 30]. Таким образом, энтропия должна была отражать сложность объекта. Далее, рассуждает Колмогоров, сложность числа, сообщения или набора данных есть величина, противоположная простоте и порядку, и она соответствует информации. Чем сложнее объект, тем больше он несет информации и – наоборот. При этом сложность объекта может быть определена через размер наименьшей компьютерной программы, необходимой для его создания. Что касается самой информации, то она, по мнению Колмогорова, является абстракцией, эквивалентной таким абстракциям, как случайность и сложность, всегда связанным друг с другом как тайные любовники. К этому добавим, что сегодня, по мнению ряда ученых, «колмогровская сложность» лежит в основе эффекта Матфея[5].

В завершение обзора развития теоретических представлений об информации Глик обращается к квантовой теории информации, возникшей в конце XX века на стыке квантовой механики, теории алгоритмов и теории информации. Ее основное назначение связано с пониманием квантового микромира. Сущность квантовой теории информации
была сформулирована американским ученым, открывшим «черные дыры» и «кротовые норы» Джоном Арчибальдом Уилером (1911-2008 гг.). В 1989 г., он заявил, что все состоит из бита (It from bit), что все физические сущности являются информационно-теоретическими в своей основе, поскольку в их основании лежит бит информации как мельчайшая и неделимая ее частица. Из этого следует, что на квантовом уровне информация неуничтожима и носит всеобщий характер. Иллюстрируя это положение, директор Института квантовой информации в Калифорнийском политехническом институте Джон Прескилл, заявил, что даже когда сгорает книга, если, в терминах физиков, вы можете проследить каждый фотон, каждую частицу пепла, то существует возможность ее собрать (с. 382). Определенные доказательства справедливости квантового свойства информации были получены английским физиком Стивеном Хокингом в ходе его изучения космических «черных дыр». Изучая их в первой половине 1970-х гг. он высказал гипотезу о том, что «черные дыры» не только поглощают материю, содержащую в своей структуре и квантовых состояниях информацию о самой себе, но и испаряются (излучение Хокинга), выделяя тепловую энергию, не несущую никакую информацию. Тогда возникает вопрос о том, что происходит с информацией, которая согласно квантовой механике не может быть уничтожена и должна сохраняться? Однако уже в 2004 г. Хокинг меняет свое мнение по поводу исчезновения информации в открытом им излучении. Информация, констатирует он, все же покидает «черную дыру», тем самым подтверждая ее неуничтожимость на квантовом уровне.

Как же понимается информация в рамках квантовой теории? Во-первых, она не просто абстрактное понятие, подобно колмогоровской сложности, она – физическая величина, и, следовательно, может быть локализована. Во-вторых, она связывается с таким понятием, как запутанность, при котором два или более квантовых объекта оказываются взаимозависимыми. Одним из приложений квантовой теории информации являются квантовые вычисления и квантовые компьютеры.

Подводя итоги рассмотрения этого раздела «Информации», можно отметить, что Глик акцентировал внимание преимущественно на наиболее ярких и известных теоретических подходах физико-математической направленности. Вполне логично, учитывая тот факт, что автор американец, акцент им был сделан на англоязычных исследованиях. Как нам представляется, целью автора было стремление продемонстрировать неоднозначность информации, как научного понятия, в силу чего вопрос о том, что такое информация должен быть заменен вопросом о том, как нам следует ее понимать. В этой ситуации трудно не согласиться с Н. Н. Моисеевым, который утверждал, что « … строгого и достаточно универсального определения информации не только нет, но и быть не может» [10, с. 106]. Это, во-первых. Во-вторых, несмотря на сложность информации как научного понятия, его теоретическая разработка способствовала созданию все более совершенных информационных технологий, ежедневно обрушивающих на нас новые, более мощные потоки информации.

[5] Эффект Матфея – это феномен неравномерного распределения преимуществ, в котором сторона, уже ими обладающая, продолжает их накапливать и приумножать, в то время как другая, изначально ограниченная, оказывается обделена ещё сильнее и, следовательно, имеет меньшие шансы на дальнейший успех.

ЗАКЛЮЧЕНИЕ

Изобретая все новые средства коммуникации, разрабатывая информационные теории, современный человек постепенно создал вокруг себя поток нескончаемой и неуничтожимой информации, приблизительным объемом 10 в 90 степени бит. Одним из первых это почувствовал английский поэт Александр Поуп (1688-1744 гг.) В те дни, когда, (после того как провидение, позволило изобретение печати как кару за грехи образованных), бумага станет столь дешевой, а печатные станки столь многочисленными, на Земле случится потоп авторов (с. 428). Именно метафора потопа представляется Глику наиболее удачной для характеристики информационного потока, не только несущего нам знания, но и грозящего «потопить» нас. Вместе с теорией информации, появились и «перегрузка информацией», и «избыток информации», и «информационная тревожность», и «информационная усталость» (с. 429). Информация, как в свое время предсказал Шеннон, начинает превращаться в энтропию данных, разрушая знания, сея хаос всезнайства. Вал данных слишком часто не дает того, что надо знать. К сожалению, цитирует он американского философа и историка Льюиса Мамфорда (1895–1990 гг.), «поиск информации», несмотря на скорость, с которой он происходит, не заменяет знаний, полученных непосредственным личным изучением, и прослеживания их пути в собственном темпе по разветвлениям соответствующей литературы (с. 430). Что же внутри этого потока данных, грозящего захлестнуть всех и вся? Глик снова прибегает к метафоре, к мифической Вавилонской библиотеке, описанной в одноименном рассказе Х. Л. Борхесом (1899–1986). В этой библиотеке содержатся все книги на всех языках, жалкие книги и книги-пророчества, Евангелие и комментарии к этому Евангелию, и комментарий к комментарию Евангелия, историю будущего в мельчайших деталях, интерполяции каждой книги во все другие книги, истинный каталог библиотеки и бесчисленное множество фальшивых каталогов. В этой библиотеке (которую другие называют вселенной) бережно хранится вся информация. Тем не менее, в ней нельзя найти знания, и именно потому, что все истинное знание находится в ней, размещенное на полках бок о бок с ложным. В зеркальных галереях, на бесчисленных полках можно найти все и ничего. Более совершенного примера, пишет Глик, перенасыщения информацией быть не может (с. 398). Старые способы организации знаний не работают, нужны новые. Одним из таких, по мнению автора, выступает онлайновая энциклопедия «Википедия»[6], содержащая более 35 млн статей в более чем 100 языковых разделах. На русском языке, по данным на 1 января 2015 г., написано более 1 млн материалов [11]. Являясь наследницей великой Вавилонской библиотеки Борхеса, «всемирного мозга», описанного Гербертом Уэллсом в одноименном романе, она нацелена на сбор всех задокументированных знаний, отделенных от конкретных людей. «Спасение» человека видится в обуздании информации, в ее подчинении знаниям. При этом инструмент «спасения» заложен в нас самих, это наше внимание. Когда информация дешевеет, должно дорожать внимание!

В конце своего виртуального путешествия Джеймс Глик предвосхищает перерождение человека. Теперь все мы, пишет он, постоянные посетители Вавилонской библиотеки, и мы же ее библиотекари. Мы идем по коридорам, обшаривая полки и переставляя на них книги в поисках смысла среди какофонии и бессвязности, читая историю прошлого и будущего, собирая наши мысли и мысли других, и время от времени смотрим в зеркала, в которых мы можем узнать людей – тех, кого породила информация.

[6] Заместитель руководителя Рособрнадзора А. Бисеров предложил запретить в России онлайн-энциклопедию "Википедия". По его словам, ресурс не помогает, а только вредит образованию. Правда потом это было названо шуткой, однако как мы знаем, «в каждой шутке есть только доля шутки» (см. [11]).

СПИСОК ЛИТЕРАТУРЫ

1. James Gleik – Royal Society Winton Prize for Science book. – URL: https://royalsociety.org/ awards/ science-books/james-gleick. Дата обращения 29.12.2014.
2. Hartley R.V.L. Transmission information // Bell System Technical Journal. – 1928. – № 7 (July). – P. 535–563; Хартли Р.В.Л. Передача информации // Теория информации и ее приложение: сб. переводов / под ред. А. А. Харкевича. – М. : Гос. изд-во физико-математической лит-ры, 1959. – С. 5-35.
3. Shannon C. E. A mathematical theory of communication // The Bell System Technical Journal. – 1948. – Vol. 27 (July, October). – P. 379–423. 623–656; Шеннон К. Э. Математическая теория связи // Работы по теории информации и кибернетике: [сборник статей] / пер. с англ. – М.: Изд-во иностранной лит-ры, 1963. – С. 243-332.
4. Урсул А.Д. Природа информации.: философский очерк. – М. : Полит. лит-ра, 1968. – 288 с.
5. Shannon C. E., Weaver W. The mathematical theory of communication. – Urbana: The University of Illinois Press, 1964.
6. Wienner N. Cybernetics or control and communication in the animal and the machine. The Technology press and John Wiley & Soris Inc. New York – Herman et Cie, Paris, 1948. – 99 p.
7. Solomonoff R.J. A preliminary report on a general theory of inductive inference, Tech. Rept. ZTB-138, Zator Company, Cambridge, Mass., November 1960.
8. Колмогоров А.Н. Три подхода к определению понятия «количество информации» // Проблемы передачи информации. – 1965. – Т.1, №1. – С. 3-11.
9. Колмогоров А. Н. Алгоритм, информация, сложность. – М.: Знание, 1991. – 48 с.

10. Моисеев Н.Н. Современный рационализм. – М. : МГВП КОКС, 1995. – 376 с.
11. Рособрнадзор пошутил о запрете «Википедии» // Ytro.ru: ежедневная электронная газета. – URL: http://www.utro.ru/articles/2015/01/22/1230623.shtml. (Дата обращения 31.01.2015).

Материал поступил в редакцию 04.02.15.

Сведения об авторе

ПЛЕШКЕВИЧ Евгений Александрович – доктор педагогических наук, ведущий научный сотрудник отдела библиотековедения Российской государственной библиотеки, Москва.
E-mail: eap1966eap@mail.ru

Федеральное государственное бюджетное учреждение науки
ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ РОССИЙСКОЙ АКАДЕМИИ НАУК

предлагает научным работникам, аспирантам и другим специалистам в области естественных, точных и технических наук, желающим быстро и эффективно опубликовать результаты своей научной и научно-производственной деятельности, использовать способ публикации своих работ через *систему депонирования.*

«Депонирование (передача на хранение) – особый метод публикации научных работ (отдельных статей, обзоров, монографий, сборников научных трудов, материалов научных конференций, симпозиумов, съездов, семинаров) узкоспециального профиля, разрешенных в установленном порядке к открытому опубликованию, широкое тиражирование которых, как правило, в силу их узкой специализации, не считается целесообразным, а также работ широкого профиля, срочная информация о которых необходима для утверждения их приоритета. Депонирование предусматривает прием, учет, регистрацию, хранение научных работ и обязательное размещение информации о них в специальных информационных изданиях».

Подготовка и передача на депонирование научных работ происходит в соответствии с «Инструкцией о порядке депонирования научных работ по естественным, техническим, социальным и гуманитарным наукам» (М., 2013).

Депонированные научные работы находятся на хранении в депозитарии ВИНИТИ РАН, копии работ предоставляются заинтересованным организациям и специалистам на бумажном и электронном носителях и являются официальной публикацией.

Информация о депонированных научных работах включается в информационные издания ВИНИТИ РАН, в РЖ ВИНИТИ РАН и БД ВИНИТИ РАН и аннотированный библиографический указатель «Депонированные научные работы».

Подать научную работу на депонирование можно, обратившись в Отдел депонирования ВИНИТИ РАН по адресу:

125190, Москва, ул. Усиевича, 20.
ВИНИТИ РАН, Отдел депонирования научных работ.
Тел.: 8 (499) 155-43-28, Факс: 8 (499) 943-00-60.
e-mail: dep@viniti.ru

С инструкцией о порядке депонирования можно ознакомиться на сайте ВИНИТИ РАН: **http://www.viniti.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: 8 (499) 155-46-20

Телефон/Факс: 8 (499) 155-45-25

E-mail: zinovyeva@viniti.ru, davydova@viniti.ru