

С.Н. Тронин

О СООТНОШЕНИИ МЕЖДУ МОЩНОСТЯМИ 1-ОРБИТ И 1 – ИНФОРМАЦИЕЙ СЛОВ

В данной заметке уточняются результаты главы 5 книги [1] (теоремы 5.1 и 5.2), доказательства которых в [1] основаны на одном неточно сформулированном (и потому фактически не доказанном) утверждении о связи между словами и эйлеровыми графами ([1], с. 77 – 78). Изложение в [1] основано, по-видимому, на работе [2] (с. 107 – 108), где тоже есть неточности, причем даже некоторые правильные утверждения из [2] почему-то преобразились, оказавшись в [1]. Мы приводим точные формулировки и подробные доказательства, так как только чрезмерной сжатостью изложения можно объяснить то, что имеющиеся пробелы оставались более 10 лет незамеченными.

Напомним основные определения и обозначения. Рассматривается конечное множество (алфавит) X , и пусть $q = |X|$ есть число его элементов. Элементы X^n , то есть упорядоченные последовательности $x = x_1x_2 \dots x_n$, называются словами в алфавите X . На множестве X^n действует слева группа подстановок n -й степени S_n , которую, следуя [1], будем обозначать через G . Действие таково: если $\sigma \in G$, то $\sigma x = x_{\sigma^{-1}(1)}x_{\sigma^{-1}(2)} \dots x_{\sigma^{-1}(n)}$. Стабилизатор слова x обозначается через G_x , это множество всех таких $\sigma \in G$, что $\sigma x = x$. Орбита Gx слова x (множество слов, образованных путем всевозможных перестановок букв в слове x) полностью определяется упорядоченной последовательностью целых неотрицательных чисел (m_1, \dots, m_q) , в которой m_i есть число вхождений i -й буквы X в слово x . При этом количество элементов орбиты $|Gx|$ равно индексу $|G : G_x|$ подгруппы G_x в группе G , и вычисляется по формуле $|Gx| = \frac{n!}{m_1!m_2! \dots m_q!}$. Последовательность (m_1, \dots, m_q) называется композицией слова x . 0-информацией слова x называется величина $I_0(x) = \log |G : G_x| = \log |Gx|$. Условная 0-информация слова $x \in X^n$ при условии $y \in X^n$ есть $I_0(x/y) = \log |G_y : (G_x \cap G_y)| = \log |G_y x|$. Обозначим через T ,

$T \in G$, циклическую перестановку букв в слове x , то есть $Tx = x_2x_3 \dots x_nx_1$. Тогда 1-информация слова x по определению есть величина $I_1(x) = I_0(x/Tx)$. Паре слов $x, y \in X^n$, $x = x_1 \dots x_n$, $y = y_1 \dots y_n$, соответствует слово $(x_1, y_1)(x_2, y_2) \dots (x_n, y_n)$ в алфавите $X \times X$, которое мы будем обозначать через $x \circ y$ (в [1] используется значок тензорного произведения \otimes , что, по нашему мнению, нецелесообразно). Орбита слова x называется в [1] также 0-орбитой. Это — множество слов x' с той же композицией, что и у слова x . По аналогии с этим 1-орбита x определяется как множество слов x' , у которых композиция $x' \circ Tx'$ совпадает с композицией слова $x \circ Tx$ (это слова в алфавите $X \times X$). Будем обозначать 1-орбиту слова x через $O_x^{(1)}$.

Основными результатами заметки являются следующие теоремы, являющиеся исправленными версиями теорем 5.1 и 5.2 из [1].

Теорема 1. Пусть t есть число различных букв в слове x , $q = |X|$. Справедливы следующие неравенства:

$$I_1(x) - (q - 1) \log n + \log t \leq \log |O_x^{(1)}| \leq I_1(x) + \log t$$

$$I_1(x) - \sum_{i=1}^q \log m_i + \log t \leq \log |O_x^{(1)}| \leq I_1(x) + \log t$$

Теорема 2. Существует префиксный код с длинами слов

$$l(x) =]I_1(x)[+] \log q[+]q^2 \log n[$$

Для доказательства теоремы 1 необходимо установить связь между словами и эйлеровыми графами. Слово $x \circ y$ интерпретируется как форма задания ориентированного графа с множеством вершин X , дугами которого являются все упорядоченные пары (x_i, y_i) , и только они ("стрелки" из x_i в y_i , $1 \leq i \leq n$, с учётом кратностей). В [1] также используется представление $x \circ y$ в виде таблицы (или матрицы) из двух строк x и y , что позволяет говорить о парах (или дугах соответствующего графа) (x_i, y_i) как о столбцах этой матрицы вида $\begin{pmatrix} x_i \\ y_i \end{pmatrix}$. Парам слов вида $x \circ Tx$ соответ-

ствуют ориентированные эйлеровы графы. 1-орбита слова x — это множество тех x' , для которых $x' \circ Tx'$ задает тот же граф, что и $x \circ Tx$.

Будем считать, что на алфавите X задан и *заксирован* некоторый линейный порядок. Положим $E = \{x \circ Tx \mid x \in X^n\}$, и пусть W есть множество пар $(x_0 \circ y_0, a)$, где $x_0, y_0 \in X^n$, a — буква слова x_0 , буквы в x_0 упорядочены относительно зафиксированного на X порядка, композиции слов x_0 и y_0 совпадают, и, если для каждой буквы $b \neq a$ из x_0 в множестве всех столбцов вида $\begin{pmatrix} b \\ b' \end{pmatrix}$ из $x_0 \circ y_0$ выбрать самый правый столбец, то совокупность таких выбранных столбцов — дуг ориентированного графа образует остовное дерево, входящее в a . Под остовным деревом, входящим в вершину a , понимается ориентированный подграф, в котором для каждой вершины b существует *единственный* маршрут из b в a . Легко показывается, что в ориентированном эйлеровом графе существуют остовные деревья, входящие в каждую вершину графа.

Лемма. *Существует биекция $\varphi : E \rightarrow W$, $\psi = \varphi^{-1}$.*

Доказательство леммы. Пусть $x \circ Tx \in E$, и a — первая (то есть самая левая) буква слова x . Тогда $\varphi(x \circ Tx) = (x_0 \circ y_0, a)$ строится следующим образом. $x_0 \circ y_0$ получается из $x \circ Tx$ перестановкой столбцов, причем должны выполняться следующие два условия: если столбец $\begin{pmatrix} b \\ b' \end{pmatrix}$ находился левее столбца $\begin{pmatrix} c \\ c' \end{pmatrix}$ в $x \circ Tx$, то он будет располагаться левее его и в $x_0 \circ y_0$, и кроме того, буквы в слове x_0 упорядочены относительно заданного на X порядка. Легко увидеть, что эти условия задают $x_0 \circ y_0$ однозначно. Из построения также видно, что пары слов $x \circ Tx$ и $x_0 \circ y_0$ определяют один и тот же граф. Далее необходимо убедиться, что $(x_0 \circ y_0, a) \in W$. Равенство композиций очевидно, так что все сводится к проверке того, что для каждой буквы $b \neq a$ из x_0 совокупность самых правых столбцов вида $\begin{pmatrix} b \\ b' \end{pmatrix}$ из $x_0 \circ y_0$ образует остовное дерево, входящее в a . По построению, самый правый такой столбец $\begin{pmatrix} b \\ b' \end{pmatrix}$ будет также самым правым и в $x \circ Tx$ (то есть любой другой столбец вида $\begin{pmatrix} b \\ b'' \end{pmatrix}$ будет располагаться левее). Пусть w — та из букв, входящих в x и не равных a , правее самого правого вхождения которой в x нет других букв, кроме,

может быть, буквы a . (Случай, когда все буквы в x равны a , тривиален, основное дерево пусто). Утверждается, что в самом правом (и в $x_0 \circ y_0$, и в $x \circ Tx$) столбце вида $\begin{pmatrix} w \\ w' \end{pmatrix}$ обязательно $w' = a$. Если w — последняя буква в слове x , то по определению Tx должно быть $x \circ Tx = \begin{pmatrix} a \dots vw \\ a' \dots wa \end{pmatrix}$. Если же w — не последняя буква в слове x , то $x \circ Tx = \begin{pmatrix} a \dots ww' \dots \\ a' \dots w' \dots \end{pmatrix}$, и, по выбору w , должно быть $w' = a$. Рассмотрим подграф D графа, задаваемого $x_0 \circ y_0$, состоящий из самых правых столбцов вида $\begin{pmatrix} b \\ b' \end{pmatrix}$, для всех букв $b \neq a$. Как только что показано, a есть вершина этого графа, существует дуга (стрелка), входящая в a , и, по построению, для каждой другой вершины b существует ровно одна стрелка (столбец), выходящая из b . Утверждается, что для любой такой вершины b существует, притом ровно один, маршрут из b в a по дугам (стрелкам) из D . Обратимся снова к слову $x \circ Tx = \begin{pmatrix} \dots bb' \dots cc' \dots ww \dots \\ \dots b' \dots c' \dots a \dots \end{pmatrix}$. Здесь $b' \neq b$, так как b' расположен в верхней строке правее b . Если $b' = a$, то маршрут завершен, так как не существует столбцов из D с буквой a в первой строке. Если же $b' \neq a$, то $b' = c$ для некоторой буквы c , причем самый правый столбец вида $\begin{pmatrix} c \\ c' \end{pmatrix}$ расположен в $x \circ Tx$ правее столбца $\begin{pmatrix} b \\ c \end{pmatrix}$. Получаем единственно возможную дугу, продолжающую маршрут, начавшийся в вершине b , причем эта дуга располагается в $x \circ Tx$ строго правее. Ввиду конечности длины слов через некоторое число шагов будет достигнута дуга, входящая в вершину a .

Обратное отображение $\psi : W \rightarrow E$ строится следующим образом (алгоритм "вычеркивания"). Пусть дан элемент $(x_0 \circ y_0, a) \in W$. Положим $t = \emptyset$ (это "текущее значение" строящегося результата, которое в конце концов окажется равным $x \circ Tx$), $z = x_0 \circ y_0$. Сначала берется самый левый столбец $\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ в z , такой, что $a_1 = a$. Он приписывается справа к слову t , и "вычеркивается" из слова z (результат "вычеркивания" снова обозначаем через z). Далее, для произвольного i , если на предыдущем шаге был "вычеркнут" из z столбец $\begin{pmatrix} a_{i-1} \\ a_i \end{pmatrix}$, то выбираем в z самый левый столбец вида $\begin{pmatrix} a_i \\ a_{i+1} \end{pmatrix}$ (то есть самый левый среди всех с буквой a_i в первой строке), приписываем его справа к t , и "вычеркиваем" из z . Таким образом, строящееся слово t будет иметь вид: $\begin{pmatrix} a_1 a_2 \dots a_{i-1} a_i & \dots \\ a_2 a_3 \dots a_i & a_{i+1} \dots \end{pmatrix}$, причем $a_1 = a$, и если будет доказано, что будут

”вычеркнуты” все столбцы z , то в результате получится $x \circ Tx$ для некоторого x . Препятствием для завершения работы алгоритма может стать ситуация, когда вычеркнут столбец $\begin{pmatrix} b \\ c \end{pmatrix}$, но в z уже не осталось буквы c в первой строке. Это возможно только в том случае, если все столбцы вида $\begin{pmatrix} c \\ d \end{pmatrix}$ уже были вычеркнуты ранее. Пусть $c \neq a$. Из определения алгоритма следует, что до столбца $\begin{pmatrix} c \\ d \end{pmatrix}$ должен быть вычеркнут столбец вида $\begin{pmatrix} s \\ c \end{pmatrix}$. Но, так как и в первой, и во второй строке содержится одинаковое число вхождений буквы c , то буква c в $\begin{pmatrix} b \\ c \end{pmatrix}$ оказывается ”лишней”. Если же $c = a$, то этого противоречия не возникнет, так как вычеркивание начинается со столбца с буквой a в первой строке. Если алгоритм не может продолжать работу после вычеркивания столбца $\begin{pmatrix} b \\ a \end{pmatrix}$, то все столбцы, содержащие букву a , уже вычеркнуты. Утверждается, что вычеркнуты все столбцы, то есть работа алгоритма завершилась успешно. Если бы это было не так, то не вычеркнутой останется и часть столбцов, входящих в остовное дерево D , так как если какой-то столбец вида $\begin{pmatrix} b \\ c \end{pmatrix}$ вычеркнут, то прежде должны быть вычеркнуты и все столбцы вида $\begin{pmatrix} b \\ b' \end{pmatrix}$, расположенные левее его. При этом, когда процесс вычеркивания заканчивается (невозможно его продолжать), для каждой буквы b количество ее вхождений в первую и вторую строки уменьшается на одну и ту же величину. Следовательно, для оставшегося слова z количество вхождений каждой буквы в верхнюю и нижнюю строки z одинаково. Пусть имеется невычеркнутый $\begin{pmatrix} b \\ c \end{pmatrix} \in D$. Тогда, чтобы баланс вхождений букв в первую и вторую строки был соблюден, необходимо наличие невычеркнутого $\begin{pmatrix} c \\ d \end{pmatrix} \in D$ (при этом $b \neq c$). То же соображение применимо к $\begin{pmatrix} c \\ d \end{pmatrix}$. Так как D — остовное дерево, то циклы невозможны, и получаем путь в остовном дереве, для продолжения которого нет формальных препятствий, кроме его окончания в вершине a . Но все столбцы, в которых содержалась буква a , уже вычеркнуты. Это показывает, что алгоритм вычеркивания успешно завершается, причем последним вычеркивается столбец вида $\begin{pmatrix} b \\ a \end{pmatrix}$.

Покажем, что $\psi\varphi = 1$. Пусть $\varphi(x \circ Tx) = (x_0 \circ y_0, a_1)$. Перенумеруем столбцы в $x \circ Tx = \begin{pmatrix} a_1 a_2 \dots a_i & \dots a_n \\ a_2 a_3 \dots a_{i+1} \dots a_1 \end{pmatrix}$, так что $\begin{pmatrix} a_i \\ a_{i+1} \end{pmatrix}$ имеет номер i , $\begin{pmatrix} a_n \\ a_1 \end{pmatrix}$ — номер n . Дополним

определение φ , предполагая, что для каждого столбца $x_0 \circ y_0$ "запоминается" номер, который был у этого столбца в $x \circ Tx$. Будем называть его меткой столбца. Таким образом, столбец имеет метку j в $x_0 \circ y_0$ тогда и только тогда, если j есть его номер в $x \circ Tx$. Индукцией по j покажем, что вычеркиваемый на j -м шаге алгоритма столбец $x_0 \circ y_0$ имеет метку j для всех $j = 1, 2, \dots, n$ (и, таким образом, $\psi(\varphi(x \circ Tx)) = x \circ Tx$). При $j = 1$ это следует из определений. Допустим, что для $1, \dots, j - 1$ утверждение доказано, и рассмотрим столбец $\begin{pmatrix} b \\ c \end{pmatrix}$, имеющий в $x \circ Tx$ номер $j - 1$, и столбец $\begin{pmatrix} c \\ d \end{pmatrix}$ с номером j . По предположению, $\begin{pmatrix} b \\ c \end{pmatrix}$ вычеркнут из $x_0 \circ y_0$ на $j - 1$ -м шаге. Выясним, где расположен $\begin{pmatrix} c \\ d \end{pmatrix}$ в $x_0 \circ y_0$. Если $c = b$, то, по определению φ , столбец $\begin{pmatrix} b \\ d \end{pmatrix}$ стоит в $x_0 \circ y_0$ сразу после $\begin{pmatrix} b \\ b \end{pmatrix}$ (то есть справа от него), и должен быть вычеркнут сразу после него, то есть на j -м шаге. Если же $c \neq b$, то все равно, по определению φ , столбец $\begin{pmatrix} c \\ d \end{pmatrix}$ будет расположен в $x_0 \circ y_0$ правее самого правого столбца вида $\begin{pmatrix} c \\ c \end{pmatrix}$, который был левее $\begin{pmatrix} b \\ c \end{pmatrix}$ в $x \circ Tx$. Номер этого столбца (а значит, и его метка) строго меньше $j - 1$, поэтому к моменту, когда на $j - 1$ -м шаге вычеркнут $\begin{pmatrix} b \\ c \end{pmatrix}$, уже вычеркнут и $\begin{pmatrix} c \\ d \end{pmatrix}$. Это означает, что $\begin{pmatrix} c \\ d \end{pmatrix}$ оказывается самым левым из невычеркнутых из $x_0 \circ y_0$ столбцов с буквой c в первой строке. Следовательно, он должен быть вычеркнут сразу после $\begin{pmatrix} b \\ c \end{pmatrix}$, то есть на j -м шаге.

Чтобы убедиться, что $\varphi\psi(x_0 \circ y_0, a) = (x_0 \circ y_0, a)$, достаточно заметить, что $(x_0 \circ y_0, a)$ полностью определяется упорядочением x_0 , буквой a , взаимным расположением столбцов, в первых строках которых стоят одинаковые буквы, и что оба отображения φ и ψ сохраняют последние два свойства. Лемма доказана.

Отличие этой леммы от того утверждения, которое использовано в [1], состоит в определении множества W . На самом деле, как только что показано, вместе со словом $x_0 \circ y_0$ необходимо указывать и букву a . Поэтому количество элементов в E в $q = |X|$ раз больше, чем это имелось в виду в книге [1]. Что касается работы [2], то в ней есть признаки того, что используется правильное определение W , но приведенное на с. 107 неравенство (аналогом которого является наша теорема 1), снова заставляет в этом усомниться.

Доказательство теоремы 1. Без ограничения общности можно считать, что $X = \{a_1, \dots, a_q\}$, причем a_1, \dots, a_t — все те буквы, которые входят в данное слово x . Будем также предполагать, что 1-орбита есть подмножество множества E , проводя отождествление с помощью биекции $X^n \rightarrow E$, $y \mapsto y \circ T y$. Пусть $E = \bigcup_{i=1}^q E_i$, где E_i — множество тех $y \circ T y$, для которых a_i есть крайняя левая буква слова y . Аналогично, пусть $W_i = \{(x_0 \circ y_0, a_i) | (x_0 \circ y_0, a_i) \in W\}$. Очевидно, что $E_i \cap E_j = \emptyset$, $W_i \cap W_j = \emptyset$ при $i \neq j$, $\psi(E_i) = W_i$, $\varphi(W_i) = E_i$ для всех i . Далее, $O_x^{(1)} = \bigcup_{i=1}^q O_x^{(1)} \cap E_i$. Положим $Y_i = O_x^{(1)} \cap E_i$. Это — непересекающиеся множества, и, по предположению, непустыми являются только Y_1, \dots, Y_t , так что $|O_x^{(1)}| = \sum_{i=1}^t |Y_i|$. Предположим для определенности, что $|Y_1| = \max_{1 \leq i \leq t} |Y_i|$, $|Y_t| = \min_{1 \leq i \leq t} |Y_i|$. Тогда $t|Y_t| \leq |O_x^{(1)}| \leq t|Y_1|$, откуда получаем $\log t + \log |Y_t| \leq \log |O_x^{(1)}| \leq \log t + \log |Y_1|$. Оценим $|Y_1|$ сверху, а $|Y_t|$ снизу. Именно это фактически и сделано в [1], [2]. Начнем с $|Y_1|$. Зафиксируем букву a_1 (можно считать ее началом слова x). Так как ограничение на W_1 отображения $W \rightarrow (X \times X)^n$, $(x_0 \circ y_0, a_i) \mapsto x_0 \circ y_0$, инъективно, то можно, "забыв" про a_1 , рассуждать так, как будто φ отображает E_1 инъективно в $(X \times X)^n$. Пусть $\varphi(x \circ T x) = x_0 \circ y_0$. Для любого $x' \circ T x' \in E_1$ будем иметь $\varphi(x' \circ T x') = (\sigma x_0) \circ (\sigma y_0)$, где $\sigma \in S_n$ — некоторая подстановка, причем $\sigma x_0 = x_0$. Это верно по той причине, что, по определению 1-орбиты, $x \circ T x$ и $x' \circ T x'$ состоят из одних и тех же столбцов, а при действии отображения φ происходит упорядочение (перестановка столбцов) по буквам первой строки. Следовательно, имеет место инъективное отображение $x' \circ T x' \mapsto x_0 \circ \sigma y_0 \mapsto \sigma y_0$ из E_1 в "относительную орбиту" $G_{x_0 y_0}$. Отсюда $|Y_1| = |\varphi(Y_1)| \leq |G_{x_0 y_0}|$. Переходя к логарифмам, получим $\log |Y_1| \leq \log |G_{x_0 y_0}| = I_0(y_0/x_0) = I_0(Tx/x) = I_0(x/Tx) = I_1(x)$. Отсюда $\log |O_x^{(1)}| \leq I_1(x) + \log t$.

Рассмотрим Y_t , и слово $x_0 \circ \sigma y_0 \in \varphi(Y_t)$, $\sigma \in G_{x_0}$. По определению W , множество D всех столбцов, в первой строке которых находится буква a_i , отличная от a_t , и таких, что среди столбцов вида $\begin{pmatrix} a_i \\ * \end{pmatrix}$ они самые правые, образует остовное дерево графа, соответствующего $x_0 \circ \sigma y_0$ (и это тот же граф, что и для $x_0 \circ y_0$, и для $x \circ T x$). Это дерево входит в a_t . Рассмотрим подгруппу G' группы G_{x_0} , состо-

ящую из всех подстановок τ , оставляющих неподвижными столбцы из множества D (при действии $\tau(x_0 \circ \sigma y_0) = (\tau x_0) \circ (\tau \sigma y_0) = x_0 \circ (\tau \sigma y_0)$). Результат этого действия принадлежит $\varphi(Y_t)$. Во-первых, он принадлежит W_t , так как сохраняются все определяющие элементы W_t свойства, в частности — наличие остовного дерева (так как граф не меняется). Во-вторых, по той же самой причине $\psi(x_0 \circ (\tau \sigma y_0)) = x' \circ T x'$ должен принадлежать 1-орбите слова x .

Итак, нашлось подмножество $\varphi(Y_t)$, число элементов которого равно порядку группы G' , а этот порядок можно оценить снизу, точно так же, как это сделано в [1]. Получаем неравенства $I_1(x) - (q-1) \log n \leq \log |\varphi(Y_t)| = \log |Y_t|$ и $I_1(x) - \sum_{i=1}^q \log m_i \leq \log |Y_t|$.

Д о к а з а т е л ь с т в о теоремы 2, по сути, точно такое же, как и в [1], с. 85–86. Множество X^n представляется в виде объединения непересекающихся 1-орбит. Кодовое слово для x составляется из префикса, в котором записывается номер 1-орбиты, и суффикса, в котором записывается номер слова внутри 1-орбиты. Для записи префикса требуется $\lceil q^2 \log n \rceil$ бит. Для записи номера слова x внутри 1-орбиты, по теореме 1 достаточно $\lceil I_1(x) \rceil + \lceil \log t \rceil \leq \lceil I_1(x) \rceil + \lceil \log q \rceil$ бит.

Заметим еще, что возможно строить код, разбивая 1-орбиты на подмножества Y_i , так что кодовые слова будут состоять из трех частей: префикса длины $\lceil q^2 \log n \rceil$, части длины $\lceil \log q \rceil$, в которой содержится номер множества i , $1 \leq i \leq q$, и суффикса, в котором содержится номер слова x внутри Y_i . В результате получается та же оценка, что и выше.

Автор выражает благодарность рецензенту за указанную погрешность в первом варианте данной заметки.

СПИСОК ЛИТЕРАТУРЫ

1. Гоппа В.Д. Введение в алгебраическую теорию информации. — М.: Наука. Физматлит, 1995. — 112 с.
2. Гоппа В.Д. Коды и информация // Успехи мат. наук. — 1984. — Т. 39, вып. 1. — С. 77 – 120.