

На правах рукописи

Мухина Светлана Анатольевна

Диаграмма Хассе частичного порядка
“быть фрагментом”

Специальность 01.01.09 - дискретная математика и математическая кибернетика

АВТОРЕФЕРАТ

Диссертация на соискание ученой степени
кандидата физико-математических наук

Москва – 2009

Работа выполнена на кафедре математических методов прогнозирования факультета Вычислительной математики и кибернетики Московского государственного университета им. М.В. Ломоносова.

Научный руководитель: доктор физико-математических наук,
профессор В.К. Леонтьев

Официальные оппоненты: доктор физико-математических наук
Ю.Г. Сметанин
кандидат технических наук
М.М. Ланге

Ведущая организация: Московский физико-технический институт
(государственный университет)

Защита диссертации состоится “ ____ ” _____ 2009 г.

в ____ час. ____ мин. на заседании диссертационного совета Д 002.017.02 при Учреждении Российской академии наук Вычислительный центр им. А.А. Дородницына РАН по адресу: 119333, г. Москва, ул. Вавилова, д.40.

С диссертацией можно ознакомиться в библиотеке ВЦ РАН.

Автореферат разослан “ ____ ” _____ 2009 г.

Ученый секретарь диссертационного
совета Д 002.017.02 при ВЦ РАН

доктор физико-математических наук

В.В. Рязанов

Общая характеристика работы

Актуальность темы. Комбинаторная теория слов является одним из активно развивающихся разделов дискретной математики, имеющим широкие и глубокие связи с алгебраической топологией, математической лингвистикой, алгебраической теорией кодирования, математической генетикой и другими разделами математики и ее приложений. Основные проблемы комбинаторики слов связаны с изучением слов и словарных функций, а основными объектами изучения являются слова, части слова, множества слов и упорядоченность на множестве слов. Многие формулировки проблем сводятся к нахождению или описанию множеств слов, обладающих определенным перечнем свойств. В частности, сюда относятся способы описания слов, не содержащих или "избегающих" запрещенные подслова. К этому же классу можно отнести проблемы описания слов по их структурированным тем или иным образом фрагментам.

Фрагмент как часть слова несет определенную информацию о всем слове, и с этой точки зрения комбинаторные исследования на множестве слов имеют вполне содержательную теоретико-информационную основу. Исследования взаимосвязей между словами и их фрагментами актуальны и имеют практическую значимость в разных направлениях современной науки, таких как, например, исследование генетических последовательностей, теория кодирования и передачи информации с большой степенью надежности и высокой скоростью.

Изучение частичного порядка "быть фрагментом" и его диаграммы Хассе позволяет определить, структурировать и формализовать характеристики, связывающие слова и их фрагменты. Такого рода исследования позволяют решать различные задачи на множестве

слов. Во многих таких задачах фундаментальной является проблема нахождения числа слов из заданного множества, содержащих в качестве "части" фрагменты другого, тем или иным способом определенного семейства слов. Классической здесь является задача о нахождении числа слов длины n , содержащих в качестве фрагмента фиксированное слово a длины m . Не лишне отметить, что решение этой задачи не дается какой-то простой формулой и зависит от "корреляционной" функции, связанной со словом a .

В диссертации рассматриваются метрические характеристики диаграммы Хассе частичного порядка "быть фрагментом" для двоичного и q -ичного алфавитов.

Цель работы. Целью работы является исследование геометрических свойств диаграммы Хассе частичного порядка "быть фрагментом" в двоичном и q -ичном алфавитах.

Научная новизна. В работе получены точные и рекуррентные формулы, позволяющие определить мощности множеств слов, содержащих в качестве фрагмента слово с заданной композицией. Следствием этого результата являются, в частности, некоторые хорошо известные в теории информации факты, часто используемые для оценки мощности кодов, корректирующих ошибки. Кроме того, получен ряд новых результатов о диаграмме Хассе частичного порядка, определенного предикатом "быть фрагментом".

Методика исследования. Методы комбинаторного анализа.

Практическая ценность. Результаты работы могут быть применены в различных оптимизационных схемах теории кодирования, использующих отношение частичного порядка, изоморфное изученному

в диссертации.

Апробация работы. Результаты диссертации докладывались на заседании кафедры математических методов прогнозирования факультета ВМиК МГУ. По теме диссертации опубликовано 4 работы.

Объем и структура работы. Диссертация состоит из введения, двух глав и списка литературы. Общий объем – 64 страницы.

Основное содержание работы

Глава 1. Введение

Комбинаторика слов представляет собой важный раздел дискретной математики, имеющий глубокое внутреннее содержание, разнообразные связи со многими классическими областями современной математики и многочисленные приложения в теории информации, математической генетике, структурной лингвистике и т.д.

Диаграмма Хассе. Основным объектом исследования являются слова, множества слов, части слова, связь частей слова, упорядоченность на множестве слов.

Итак, пусть $B = \{a_1, a_2, \dots, a_q\}$ – q -ичный алфавит и B^* – множество всех слов конечной длины над алфавитом B . Если $a \in B^*$, то любое слово a' , полученное из a удалением некоторых букв, называется фрагментом слова a . Это отношение задает частичный порядок на B^* , который мы будем обозначать стандартным образом

$$a' \leq a, \tag{1}$$

если a' – фрагмент a .

Геометрически отношение (1) описывается с помощью диаграммы Хассе, которая представляет собой граф с множеством вершин $X =$

B^* , смежность которых определяется понятием “покрытия”. Другими словами a и b – смежны, если $a < b$ и не существует вершины z с условием $a < z < b$. Таким образом, смежными на диаграмме Хассе являются “ближайшие” сравнимые вершины.

Важнейшими понятиями, связанными с частичным порядком (1) и диаграммой Хассе, являются следующие:

- а) степени вершин;
- б) длины цепей, соединяющих вершины a и b ;
- в) мощность антицепей.

Степени вершин. Описание степеней вершин связано с разложением произвольного слова на произведение блоков или, что то же самое, с серийным представлением слов из B^* .

Определение 1. Пусть $v_i = v_{i_1}v_{i_1+1}\dots v_{i_1+r}$ – подслово (фактор) слова v . Подслово v_i называется серией v , если

- а) $v_{i_1} = v_{i_1+1} = \dots = v_{i_1+r} = a_s \in B$;
- б) $v_{i_1-1} \neq a_s, v_{i_1+r+1} \neq a_s$.

Ясно, что каждое подслово $v \in B^*$ имеет единственное представление в виде произведения (конкатенации) серий.

Длины цепей, соединяющих две вершины. Если $a, b \in B^*$ и $a < b$, то любое множество элементов

$$a \leq b_1 < b_2 < \dots < b_k \leq b \quad (2)$$

называется цепью. Цепь (2) называется максимальной или насыщенной, если в нее нельзя “вставить” ни одного элемента из B^* .

Определение 2. Длина $l(C)$ конечной цепи C определяется равенством $l(C) = |C| - 1$.

Определение 3. Если $a, b \in B^*$, то говорят, что элемент b покрывает элемент a , если $a < b$ и ни один элемент $z \in B^*$ не удовлетворяет условию $a < z < b$.

Определение 4. Элемент $a \in B^*$ называется максимальным (минимальным), если из того, что $a \leq x$ ($x \leq a$), $x \in B^*$, следует $a = x$.

Определение 5. Частично-упорядоченное множество (ч.у. множество) $P = (M, \leq)$ называется градуированным ранга n , если каждая максимальная цепь имеет одну и ту же длину n . В этом случае существует единственная ранговая функция $\rho : B^* \rightarrow \{0, 1, \dots, n\}$ такая, что $\rho(a) = 0$, если a – минимальный элемент B^* , и $\rho(b) = \rho(a) + 1$, если b покрывает a в B^* .

В нашем случае множество $(B^*, \leq)$ является градуированным, и ранговая функция слова $x \in B^*$ – это просто длина слова x , т.е. $\rho(x) = |x|$.

По определению длина интервала $[x, y]$, где $x \leq y$, – это максимальная длина цепи, принадлежащей этому интервалу, т.е. $l(x, y) = \rho(y) - \rho(x) = |y| - |x|$.

Значительный интерес для последних исследований представляет мощность интервала $[x, y]$. Если

$$\xi(x, y) = \begin{cases} 1, & x \leq y, \\ 0, & \text{если не так,} \end{cases}$$

то имеет место стандартное представление

$$\xi^2(x, y) = \sum_{x \leq z \leq y} 1 = |[x, y]| = \text{Card}[x, y],$$

где свертка $\xi^2(x, y)$ определяется стандартным способом

$$f \cdot g(x, y) = \sum_{x \leq z \leq y} f(x, z)g(z, y),$$

или в нашем случае

$$\xi^2(x, y) = \sum_{x \leq z \leq y} \xi(x, z)\xi(z, y).$$

Если опять обратиться к диаграмме Хассе, то мощность интервала $[x, y]$ при $x < y$ равна общему числу вершин, лежащих на всех путях, соединяющих x и y .

Мощность антицепей. Пусть $B = \{a_1, a_2, \dots, a_q\}$ – q -ичный алфавит с частичным порядком "быть фрагментом" ($\leq$).

Определение 6. Антицепь есть подмножество A ч.у. множества P , в котором любые два элемента несравнимы.

Определение 7. Мощность антицепи V – число элементов в ней.

Геометрические свойства диаграммы Хассе. Геометрия диаграммы Хассе частичного порядка "быть фрагментом" следующая:

- а) Если вершина x лежит на k -м слое диаграммы Хассе, т.е. $|x| = k$, то из нее "исходит" ровно $(k + 1)(q - 1) + 1$ ребер.
- б) Если вершина x лежит на k -м слое диаграммы Хассе, то в нее "входит" ровно $s(x)$ ребер, где $s(x)$ – число серий слова x .
- в) Если рассматриваются q -ичные слова длины $\leq n$, то максимальная длина антицепи в этом множестве есть q^n и эта единственная антицепь совпадает с верхним слоем диаграммы Хассе.

Глава 2. Бинарный алфавит

В главе 2 рассматриваются бинарные слова с определенными характеристиками, а именно: слова конечной длины с заданным

весом Хэмминга. Рассматривая частичный порядок “быть фрагментом”, мы находим для указанных слов соотношения для числа слов, содержащих слово в качестве фрагмента. Для найденного соотношения устанавливается связь с ранее известными фактами, а также предоставляется ряд следствий.

Хорошо известным фактом является то, что число слов конечной длины n , содержащих в качестве фрагмента данное слово a , не зависит от вида слова и определяется лишь его длиной. Формально это утверждение выглядит следующим образом:

$$F_n(a) = \sum_{i=|a|}^n \binom{n}{i}, \quad (3)$$

где $|a|$ – длина слова a .

В главе 2 этот факт получает дальнейшее развитие. Принимая во внимание структурный состав слов (как самих слов x , так и заданного фрагмента a), а также учитывая тот факт, что число слов, которые содержат данное слово a в качестве фрагмента, определяется только структурой фрагмента, мы находим точное соотношение для определения данного числа слов.

$$F_N(m, n, k) = \sum_{y=n-k}^{m+n-2k} \binom{y-1}{n-k-1} \binom{N-y}{N-m-n+k}, \quad (4)$$

где $F_N(m, n, k)$ – число слов длины N и веса Хэмминга m , содержащих в качестве фрагмента произвольное слово a длины n и веса Хэмминга k .

Далее, в качестве следствия доказывается утверждение, отражающее связь с ранее известным фактом (3).

В терминах диаграммы Хассе условие (3) выражает число путей из вершины a диаграммы Хассе частичного порядка “быть фрагментом” до вершин, расположенных на n -м слое диаграммы Хассе. Найденное

условие (4) позволяет среди числа таких путей понять, сколько путей идет к вершинам, имеющим определенный вес Хэмминга m .

Глава 3. q -ичный алфавит

В главе 3 частичный порядок “быть фрагментом” рассматривается в q -ичном алфавите. Как и для случая бинарного алфавита, находится число путей из вершины a диаграммы Хассе до вершин, расположенных на n -м слое диаграммы Хассе. Это число выражается в следующем утверждении.

Теорема 1. Справедливо соотношение

$$|T_a^n| = \sum_{i=k}^n \binom{n}{i} (q-1)^{n-i}, \quad (5)$$

где $k = |a|$.

Используя далее серийное представление слов, мы находим следующий параметр диаграммы Хассе – степень “захода” вершины a . Это число оказывается равным числу серий в слове a . Данное утверждение формально отображено в следующей лемме.

Лемма 1. Справедливо соотношение

$$r(a) = s(a),$$

где $s(a)$ – число серий слова a .

Найденное отношение позволяет определить число слов длины n над q -ичным алфавитом, имеющих ровно k серий. Это число определяется в следующей лемме.

Лемма 2. Число слов длины n над алфавитом B , имеющих ровно k серий, равно $q(q-1)^{k-1} \binom{n-1}{k-1}$.

Используя результаты леммы, мы далее определяем производящую функцию, математическое ожидание и дисперсию случайной величины

ξ_n , равномерно распределенной на множестве B^n , где B – q -ичный алфавит, и равной числу серий слова $x \in B^n$. Данные величины определяются следующими утверждениями.

Лемма 3. Производящая функция $F_n(z)$ случайной величины ξ_n имеет вид

$$F_n(z) = z \left(1 + \frac{(z-1)(q-1)}{q} \right)^{n-1}. \quad (6)$$

Следствие 1. Справедливы равенства:

$$M\xi_n = n - \frac{n-1}{q}, \quad D\xi_n = \frac{(n-1)(q-1)}{q^2}.$$

Если обозначить множество всех фрагментов q -ичного слова $x \in B^*$ через T_x и рассмотреть слово $x = \gamma_1^{t_1} \gamma_2^{t_2} \dots \gamma_k^{t_k}$ в его серийном представлении, то можно найти верхнюю и нижнюю границу для числа фрагментов слова x .

Лемма 4. Справедливы неравенства

$$t_1 t_2 \dots t_k < |T_x| < (1+t_1)(1+t_2) \dots (1+t_k). \quad (7)$$

В главе 3 приводится аналог утверждения для функции $F_N(m, n, k)$ – числа слов длины N и веса Хэмминга m , содержащих в качестве фрагмента произвольное слово a длины n и веса Хэмминга k , для q -ичного алфавита. В случае q -ичного алфавита аналогом веса Хэмминга слова предлагается использовать понятие композиции слова, которая определяется следующим образом:

Определение 8. Пусть $B = \{a_1, a_2, \dots, a_q\}$ и $a \in B^*$. Через $w_k(a)$ мы обозначим число вхождений буквы a_k в слово a . Тогда вектор $w(a) = (w_1, w_2, \dots, w_q)$ называется композицией слова a .

Вводя для q -ичного алфавита функцию $F_n(a, w_1, w_2, \dots, w_q)$, которая определяет число слов длины n и заданной композиции w , содержащих в качестве фрагмента слово a , мы сначала доказываем по аналогии

с бинарным алфавитом независимость данной функции от вида фрагмента a , т.е. $F_n(a, w_1, w_2, \dots, w_q) = F_n(t_1, t_2, \dots, t_q, w_1, w_2, \dots, w_q)$, где $w(a) = (t_1, t_2, \dots, t_q)$ – композиция слова a . Затем для функции $F_n(t_1, t_2, \dots, t_q, w_1, w_2, \dots, w_q)$ мы находим рекуррентное соотношение, которое отображено в следующей теореме:

Теорема 2. Справедливо рекуррентное соотношение

$$F_n(t_1, t_2, \dots, t_q, w_1, w_2, \dots, w_q) = \sum_{y=t_1} \sum_{r_2, \dots, r_q} \binom{y-1}{t_1-1, r_2, \dots, r_q} F_{n-y}(0, t_2, \dots, t_q, w_1-t_1, \dots, w_q-r_q). \quad (8)$$

Индексы суммирования $y, r_2, \dots, r_q$ в (8) изменяются в тех пределах, для которых определены полиномиальный коэффициент и функция F_{n-y} .

Используя полученные характеристики диаграммы Хассе частичного порядка "быть фрагментом" в теории антицепей, далее можно доказать теорему о мощности максимальной антицепи.

Теорема 3. Мощность максимальной антицепи ч.у. множества $(B^{\leq n}, \leq)$ равна q^n , и эта единственная антицепь совпадает с верхним слоем диаграммы Хассе.

Заключение

Диссертационная работа посвящена исследованию метрических характеристик диаграммы Хассе частичного порядка "быть фрагментом".

Основные результаты диссертации, выносимые на защиту:

- 1) Получена точная формула, определяющая число слов в бинарном алфавите, заданной длины и веса Хэмминга, содержащих в качестве фрагмента произвольное бинарное слово с заданной длиной и весом Хэмминга.

- 2) Доказана рекуррентная формула, определяющая число слов в q -ичном алфавите с известной композицией, содержащих в качестве фрагмента произвольное q -ичное слово с заданной композицией.

Благодарности

Автор выражает благодарность своему научному руководителю Владимиру Константиновичу Леонтьеву.

Публикации по теме диссертации

1. Леонтьев В.К., Мухина С.А. О фрагментах слов // Пробл. передачи информации. 2006. Т.42. №3. С.73-77.
2. Леонтьев В.К., Мухина С.А. О фрагментах слов над q -ичным алфавитом // Пробл. передачи информации. 2008. Т.44. №3. С.63-69.
3. Леонтьев В.К., Мухина С.А. О фрагментах q -ичных слов // Тезисы докладов XV международной конференции "Проблемы теоретической кибернетики" (Казань, 2-7 июня 2008 г). Под редакцией Ю.И. Журавлева. Казань: Отечество, 2008. С.74.
4. Мухина С.А. Нахождение числа слов фиксированной длины и фиксированного веса, содержащих в качестве фрагмента слово заданной длины и заданного веса // Сборник тезисов лучших дипломных работ 2005 года. М.: Издательский отдел Факультета ВМиК МГУ им. М.В. Ломоносова, 2005. С.59-60.