ОБ ИССЛЕДОВАНИИ КОМБИНАТОРНЫХ СВОЙСТВ СТРОЯ ЗНАКОВОЙ ЦЕПИ

Семенов А. В.
ОмГТУ
avs110@inbox.ru

Аннотация

В работе рассмотрены комбинаторные свойства строя знаковой цепи произвольной природы. Дается краткое описание строя, как нового математического объекта — особым образом организованного кортежа на основе данной знаковой последовательности.

Получены выражения для подсчета количества всех возможных разных строев для цепей с заданной длиной. Кроме того, представлены формулы вычисления числа строев с ограничением на мощность алфавита и числа вхождений отдельных компонентов цепи. Отмеченные формулы получены на основе гипотезы о взаимнооднозначном отображении множество разных строев на семейство всех неупорядоченных разбиений некоторого конечного множества.

Разработана процедура генерации всех различных строев для знаковых цепей с некоторой заданной длиной, мощностью алфавита и числами вхождений отдельных компонентов цепи. Данная процедура апробирована для задачи генерации всех разных строев для цепей малой длины.

Формулы для подсчета количества всех возможных разных строев для цепей с заданной длиной подтверждены результатами компьютерной апробации путем генерации этих строев.

Введение

В настоящее время анализ данных (Data Mining) получил широкое распространение в сфере информационных технологий. Однако это направление базируется преимущественно на аппарате математической статистики [2]. Гуменюком А.С. был разработан новый подход в описываемой области — формальный анализ строя знаковой цепи [1]. Приведем определение строя цепи:

Строй цепи сообщений — это кортеж (упорядоченное множество), в котором каждому компоненту данной цепи в соответствие поставлено натуральное число, причем идентичные по выбранному признаку компоненты отображены одним и тем же числом. Самый первый компонент такого кортежа — единица, а все остальные первые встречные разные натуральные числа (представляющие вместе с единицей алфавит строя) возрастают на единицу (см. Рис. 1).

Цепь событий	A	A	G	C	T	T	A	G	G	T
Строй цепи	1	1	2	3	4	4	1	2	2	4

Рис. 1. Произвольная знаковая цепь и соответствующий ей строй

В настоящее время отсутствуют работы посвященные рассмотрение строя как комбинаторного объекта. Кроме того, отсутствие алгоритма генерации множества всех возможных строев для цепей с заданной длинной и мощностью алфавита затрудняет исследование рассматриваемого формального объекта. В данной работе представлена разработка алгоритма генерации и вывод формул числа всех строев для цепей с заданными параметрами.

Формула вычисления числа всех строев для цепей с ограничением на длину и мощность алфавита

Рассмотрим семейство множеств всех строев знаковых цепей с различной длиной. На рисунке 2 представлены множества строев для цепей длиной $n = \overline{1,4}$.

m\n	1	2	3	4
1	1	11	111	1111
2	-	12	112	1112
			121	1121
			122	1211
				1221
				1122
				1212
				1222
3	-	-	123	1123
				1213
				1223
				1231
				1232
				1233
4	-	-	-	1234

Рис. 2. Семейство множеств всех строев цепей длиной $1 \dots 4$

Обозначим K_m^n число всех строев для цепей длиной n, мощность алфавита которых равна m.

Предположим о существовании взаимно-однозначного соответствия между всеми строями для цепей длиной n с алфавитом мощностью m и семейством разбиений n-элементного множества на m подмножеств. На основе гипотезы получим формулу для K_m^n по аналогии с выводом последней для чисел Стирлинга второго рода [6].

Рассмотрим частные случаи. При $m = 1$ существует единственный строй для любого n, представляющий собой последовательность n единиц, т. е. $K_1^n = 1$. При $n = m = 2$, $K_2^2 = 2$. Цепь длиной $n > 2$ с алфавитом мощностью 2 может быть представлена одним из $K_2^n = 2^{n-1} - 1$ строем. Действительно, если обозначить первый компонент цепи единицей, то существует 2^{n-1} способ обозначить оставшиеся компоненты цепи. Так как случай

обозначения всех компонентов цепи единицей противоречит условию $m = 2$, вычтем из общего числа строев K_2^n единицу.

Для произвольных m и n получим рекуррентное соотношение:

$$K_m^n = mK_m^{n-1} + K_{m-1}^{n-1} \tag{1}$$

Рассмотрим выражение (1). Если обозначить n-й компонент цепи равным m, то первые $n-1$ компонентов цепи могут быть обозначены K_{m-1}^{n-1} способами. Кроме того, существует K_m^{n-1} способов обозначить первые $n-1$ компонентов для каждого из m вариантов обозначения n-ого компонента

Выражение (1) соответствует рекуррентной формуле для вычисления числа Стирлинга второго рода $S_{n,m}$. Таким образом, можно говорить о справедливости выдвинутой ранее гипотезы.

При снятии ограничения на мощность алфавита, получим число всех строев K^n для цепи длиной n в виде

$$K^n = \sum_{i=1}^{n} K_m^n \tag{2}$$

Правая часть выражение (2) равна числу Белла [3]

$$K^n = B_n = \sum_{i=1}^{n} S_{n,i} \tag{3}$$

На рисунке 3 представлено соответствие между строями и соответствующими им разбиениями множества $\{a, b, c\}$.

111	$\{a, b, c\}$
111	$\{a, b\}\{c\}$
111	$\{a, c\}\{b\}$
111	$\{a\}\{b, c\}$
111	$\{a\}\{b\}\{c\}$

Рис. 3. Строи цепи длиной 3 и соответствующие разбиения

Существует более удобная формула для вычисления чисел Белла (4) [3]:

$$K^{n+1} = B^{n+1} = \sum_{i=0}^{n} \binom{n}{i} K^{n-i} \tag{4}$$

Алгоритм генерации множества всевозможных строев для цепей с заданными параметрами

Из определения строя цепи первым его компонентом является 1, второй компонент строя может принимать одно из двух возможных значений. Он будет единичным, если первый и второй символы цепи совпадают и его значение будет равен 2 в ином случае. Для значения третьего компонента строя возможны аналогичные варианты, в случае совпадения первого и второго символов цепи, в случае же если два первых компонента цепи

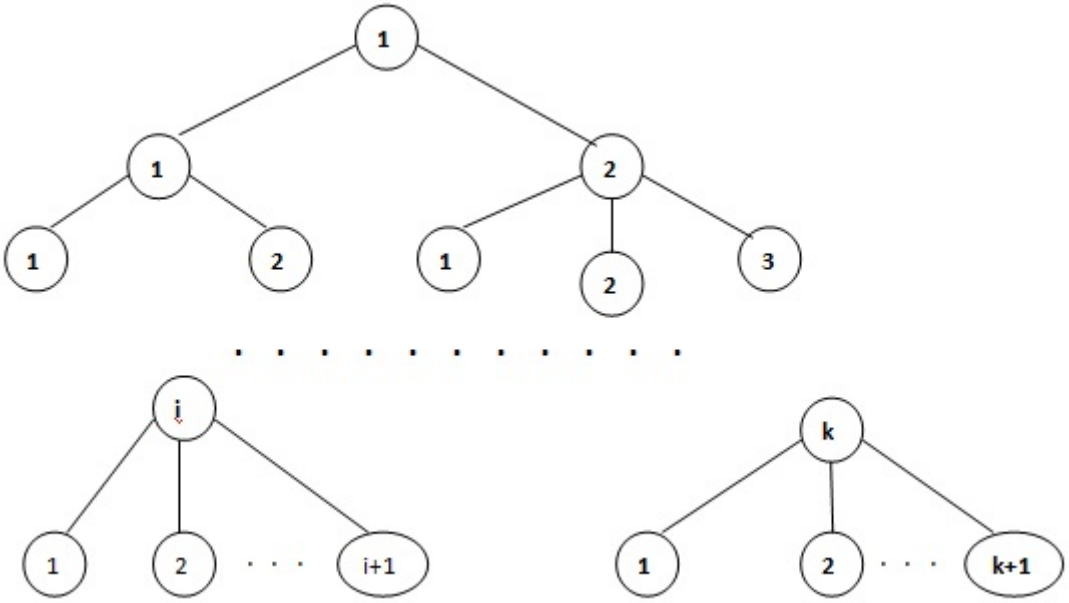

Рис. 4. Дерево-генератор множества строев

разные, третий компонент будет принимать одно из трех значений 1, 2 или 3. Таким образом, каждый последующий компонент строя примет либо одно из значений предыдущих компонентов, либо — увеличенное на единицу максимальное значение предшествующих компонентов (в случае появления нового сообщения знаковой цепи).

На основании вышесказанного всевозможные варианты строев можно представить в виде дерева [4] (Рис. 4). Между множеством путей от корня до каждого из листьев и множеством строев существует взаимно-однозначное соответствие. Каждый из узлов i-ого слоя дерева соответствует i-ому компоненту одного из строев, представленных деревом. Корень дерева соответствует первому компоненту строя, т.е. 1, листья - последнему компоненту строя. Обозначим j-й узел i-ого слоя k-ого родительского узла как $v_{i,j}^k$. Также под $v_{i,j}^k$ будем понимать значение компонента строя соответствующего узла.

Таким образом, для генерации множества возможных строев цепей длинной n и мощность алфавита m необходимо построить дерево, со следующими свойствами:

1. Количество слоев равно n;
2. $v_{1,1}^0 = 1$;
3. Количество дочерних узлов $v_{i,j}^k$ равно $\min m, v_{i,j}^k + 1$;
4. $v_{i,j}^k = \min j, m$.

Для генерации строев с помощью построенного дерева использован алгоритм обхода в прямом порядке [5].

Таким образом было получено выражение для вычисления количества всех разных строев знаковой цепи с заданными параметрами. В настоящее время алгоритм апробирован для построения всех строев с мощностью алфавита от 1 до 12 и максимальной длинной

цепи 12. В качестве примера ниже представлены некоторые короткие строи множества, сформированного автоматически:

$$(1112, 1121, 1211).$$

По мнению автора полученные выражения и процедура генерации позволят более тонко изучить свойства нового формального объекта, строя знаковой цепи.

ЛИТЕРАТУРА

[1] Гуменюк А.С., Кликушин Ю.Н., Кобенко В.Ю., Цыганенко В.Н. под науч. ред. д. т. н. Кликушина Ю.Н. Алгоритмы анализа структуры сигналов и данных: монография. — Омск: Изд-во ОмГТУ, 2010. — 272 с.

[2] Загоруйко Н. Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999.—270 с.

[3] Сачков В.Н. Введение в комбинаторные методы дискретной математики. — 2-е изд., испр. И доп. — М.: МЦНМО, 2004. — 424 с.: ил.

[4] Харари Ф. Теория графов: Пер. с англ. / Предисл. Козырева В.П.; под ред. Гаврилова Г.П. Изд. 4-е. — М.: Книжный дом «ЛИБРОКОМ», 2009. — 296 с.

[5] Кормен Т., Лейзерсон Ч., Ривест Р. Алгоритмы: построение и анализ. М. МЦНМО, 2010. — 1296 с.

[6] Грэхем Р.Л., Кнут Д.Э., Паташник О. Конкретная математика. Математические основы информатики, 2-е изд. : Пер. с англ. — М. : ООО «И.Д. Вильямс», 2010. — 784 с.