Лекция 2: Источники информации

А. М. Шур

Кафедра алгебры и фундаментальной информатики УрФУ

1 марта 2017 г.

Источник информации

На этот раз мы не будем обсуждать философские и бытовые представления, а перейдем сразу к математическому:

Источник (дискретной) информации — это

абстрактная машина M без входа, с конечным или бесконечным числом состояний и конечным выходным алфавитом Σ, работающая в дискретном времени ("тактами"). С каждым состоянием s связаны случайные величины (кубики) ξ_s и η_s. На каждом такте

- M бросает кубики, связанные с текущим состоянием s;
- значение кубика ξ_s (символ из Σ или строка над Σ) подается на выход;
- M детерминированно вычисляет состояние s' для следующего такта как функцию всех предыдущих состояний, включая s, и результатов всех состоявшихся бросков кубиков η (некоторые из аргументов могут быть фиктивными).
  - Никаких требований независимости η_s и ξ_s нет: η_s может вообще быть функцией ξ_s, в том числе совпадать с ξ_s

Источник информации

На этот раз мы не будем обсуждать философские и бытовые представления, а перейдем сразу к математическому:

Источник (дискретной) информации — это

абстрактная машина M без входа, с конечным или бесконечным числом состояний и конечным выходным алфавитом Σ, работающая в дискретном времени ("тактами"). С каждым состоянием s связаны случайные величины (кубики) ξ_s и η_s. На каждом такте

- M бросает кубики, связанные с текущим состоянием s;
- значение кубика ξ_s (символ из Σ или строка над Σ) подается на выход;
- M детерминированно вычисляет состояние s' для следующего такта как функцию всех предыдущих состояний, включая s, и результатов всех состоявшихся бросков кубиков η (некоторые из аргументов могут быть фиктивными).
  - Никаких требований независимости η_s и ξ_s нет: η_s может вообще быть функцией ξ_s, в том числе совпадать с ξ_s

Удобная для всяких приближений модель — марковские источники:

★ Источник называется марковским, если следующее состояние s' есть значение, выпавшее в текущем состоянии s на кубике η_s.

Марковские источники

Вначале выясним, причем тут

Марковские источники

Вначале выясним, причем тут

Андрей Андреевич Марков-старший.
Ему нечего терять, кроме своих цепей.

Марковские источники

Вначале выясним, причем тут

Андрей Андреевич Марков-старший.
Ему нечего терять, кроме своих цепей.

★ Дискретная марковская цепь — это случайный процесс в дискретном времени $t = 0, 1, \ldots, n, \ldots$ с дискретным множеством S состояний системы; процесс идет так:

- с каждым состоянием $s \in S$ связан кубик η_s со множеством граней S
- начальное состояние $s(0)$ задается детерминированно или определяется броском отдельного кубика, задающего начальные вероятности
- состояние $s(t+1)$ определяется броском кубика $\eta_{s(t)}$; таким образом, $s(t+1)$ зависит только от $s(t)$, но не от более ранних состояний

Марковские источники (2)

Итак, марковский источник — это марковская цепь с дополнительным функционалом (генерацией на каждом такте символов или строк в выходном алфавите).

Марковские источники (2)

Итак, марковский источник — это марковская цепь с дополнительным функционалом (генерацией на каждом такте символов или строк в выходном алфавите).

- Мы ограничимся рассмотрением марковских источников с конечным (возможно, очень большим!) числом состояний
- Марковская цепь в этом случае задается стохастической $(|S|\times|S|)$-матрицей M
  - "стохастическая" означает, что все элементы неотрицательны и сумма элементов любой строки равна 1; это позволяет интерпретировать элемент $M[i,j]$ как вероятность перехода из i-го состояния в j-е
- Каждое состояние $s \in S$ снабжается кубиком ξ_s для генерации символов/строк

Марковские источники (2)

Итак, марковский источник — это марковская цепь с дополнительным функционалом (генерацией на каждом такте символов или строк в выходном алфавите).

- Мы ограничимся рассмотрением марковских источников с конечным (возможно, очень большим!) числом состояний
- Марковская цепь в этом случае задается стохастической $(|S|\times|S|)$-матрицей M
  - "стохастическая" означает, что все элементы неотрицательны и сумма элементов любой строки равна 1; это позволяет интерпретировать элемент $M[i,j]$ как вероятность перехода из i-го состояния в j-е
- Каждое состояние $s \in S$ снабжается кубиком ξ_s для генерации символов/строк

Пример : $$M = \begin{pmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \end{pmatrix},\ \xi_1 = (a_{|0.5}, b_{|0.5}),\ \xi_2 = (a_{|0.3}, ba_{|0.5}, bbb_{|0.2})$$

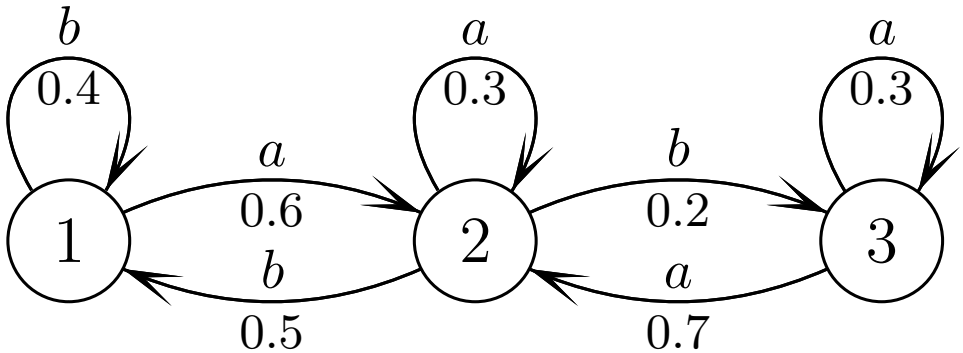
Марковские источники (2)

Итак, марковский источник — это марковская цепь с дополнительным функционалом (генерацией на каждом такте символов или строк в выходном алфавите).

- Мы ограничимся рассмотрением марковских источников с конечным (возможно, очень большим!) числом состояний
- Марковская цепь в этом случае задается стохастической $(|S|\times|S|)$-матрицей M
  - "стохастическая" означает, что все элементы неотрицательны и сумма элементов любой строки равна 1; это позволяет интерпретировать элемент $M[i,j]$ как вероятность перехода из i-го состояния в j-е
- Каждое состояние $s \in S$ снабжается кубиком ξ_s для генерации символов/строк

Пример : $M = \begin{pmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \end{pmatrix}$, $\xi_1 = (a_{|0.5}, b_{|0.5})$, $\xi_2 = (a_{|0.3}, ba_{|0.5}, bbb_{|0.2})$

Если ξ_s — функция η_s для всех $s \in S$, то источник становится вероятностным конечным автоматом:



Марковские источники для генерации текстов

За счет чего можно сжать текст (на примере текста на русском языке)?

- Символы встречаются в тексте с очень разной частотой
  - $\text{частота}(\text{о}) \approx 0.11$, $\text{частота}(\text{ъ}) \approx 0.0004$
- ★ Вероятность появления символа в данной позиции зависит от контекста
  - группы предшествующих символов (левый контекст)
  - группы последующих символов (правый контекст)
- Длина контекста, от которого зависит символ, варьируется чаще всего в диапазоне 4-6 символов
  - контекст длины k называют контекстом k-го порядка

Марковские источники для генерации текстов

За счет чего можно сжать текст (на примере текста на русском языке)?

- Символы встречаются в тексте с очень разной частотой
  - $\text{частота}(\text{о}) \approx 0.11$, $\text{частота}(\text{ъ}) \approx 0.0004$
- ★ Вероятность появления символа в данной позиции зависит от контекста
  - группы предшествующих символов (левый контекст)
  - группы последующих символов (правый контекст)
- Длина контекста, от которого зависит символ, варьируется чаще всего в диапазоне 4-6 символов
  - контекст длины k называют контекстом k-го порядка

Пример: в левом контексте медве с ненулевыми вероятностями (с точностью до опечатки) встречаются только 2 буквы: д и ж (например, медведь, медвежий)

Марковские источники для генерации текстов

За счет чего можно сжать текст (на примере текста на русском языке)?

- Символы встречаются в тексте с очень разной частотой
  - $\text{частота}(\text{о}) \approx 0.11$, $\text{частота}(\text{ъ}) \approx 0.0004$
- ★ Вероятность появления символа в данной позиции зависит от контекста
  - группы предшествующих символов (левый контекст)
  - группы последующих символов (правый контекст)
- Длина контекста, от которого зависит символ, варьируется чаще всего в диапазоне 4-6 символов
  - контекст длины k называют контекстом k-го порядка

Пример: в левом контексте медве с ненулевыми вероятностями (с точностью до опечатки) встречаются только 2 буквы: д и ж (например, медведь, медвежий)

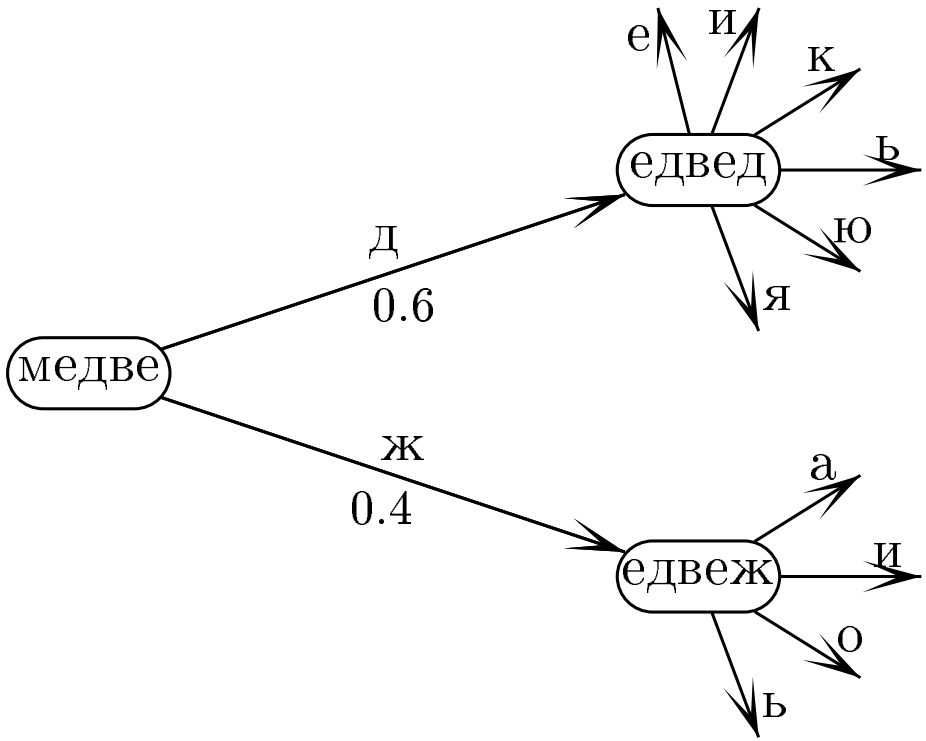
Марковский источник k-го порядка — это

дискретный марковский источник с конечным числом состояний, в котором

- $S \subseteq \Sigma^k$
- на каждом такте генерируется один символ из Σ
- если a — символ, выпавший на кубике ξ_s в состоянии $s(t) = a_1 a_2 \cdots a_k$, то $s(t+1) = a_2 \cdots a_k a$
  - таким образом, η_s — функция от ξ_s, то есть кубики η не нужны
  - при $k = 0$ получаем простейший источник, состоящий из единственного кубика ξ, как в Лекции 1

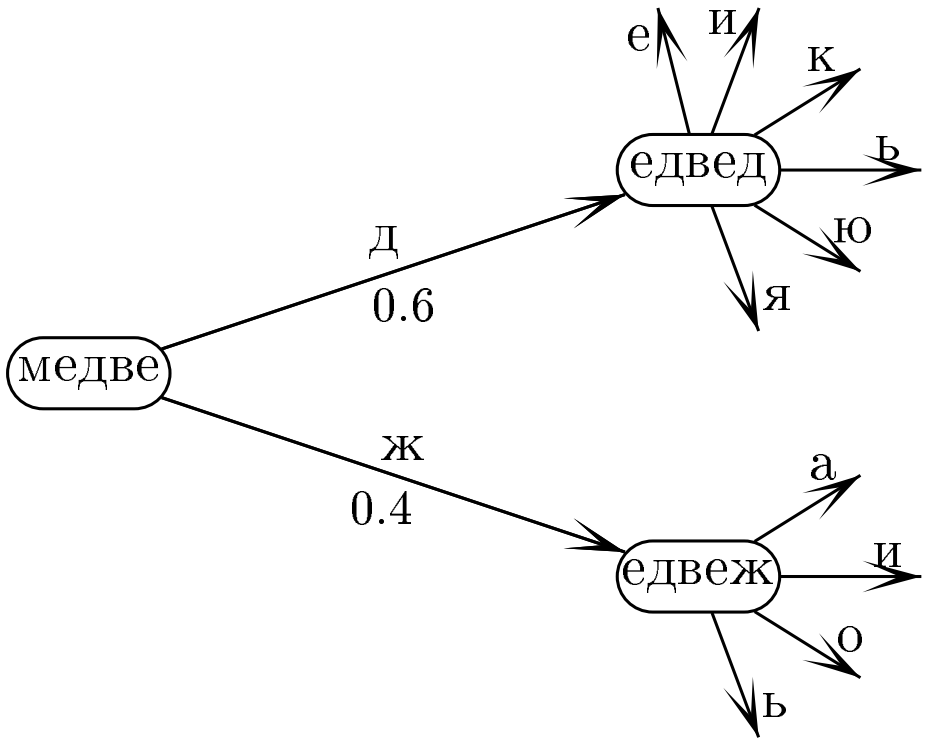
Марковские источники для генерации текстов (2)

Пример: фрагмент марковского источника 5-го порядка для русского текста



Марковские источники для генерации текстов (2)

Пример: фрагмент марковского источника 5-го порядка для русского текста



Источники могут быть и смешанного порядка — в этом случае S есть множество строк длины $\leqslant k$.

Энтропия источника

Наша цель — найти подходящий к данным источник с маленькой энтропией.

Энтропия источника

Наша цель — найти подходящий к данным источник с маленькой энтропией. Мы пока умеем считать только энтропию случайной величины, т.е. марковского источника порядка 0, например:

$\xi_1 = (x_{1|1/6}, x_{2|1/6}, x_{3|1/6}, x_{4|1/8}, x_{5|1/8}, x_{6|1/8}, x_{7|1/8})$;

$H(\xi_1) = 3 \cdot \frac{1}{6} \log 6 + 4 \cdot \frac{1}{8} \log 8 = \log \sqrt{48} = 2 + \log \sqrt{3}$.

$\xi_2 = (x_{1|1/4}, x_{2|1/4}, x_{3|1/12}, x_{4|1/12}, x_{5|1/12}, x_{6|1/12}, x_{7|1/12}, x_{8|1/12})$;

$H(\xi_2) = 2 \cdot \frac{1}{4} \log 4 + 6 \cdot \frac{1}{12} \log 12 = \log \sqrt{48} = 2 + \log \sqrt{3}$.

Энтропия источника

Наша цель — найти подходящий к данным источник с маленькой энтропией. Мы пока умеем считать только энтропию случайной величины, т.е. марковского источника порядка 0, например:

$\xi_1 = (x_{1|1/6}, x_{2|1/6}, x_{3|1/6}, x_{4|1/8}, x_{5|1/8}, x_{6|1/8}, x_{7|1/8})$;

$H(\xi_1) = 3 \cdot \frac{1}{6} \log 6 + 4 \cdot \frac{1}{8} \log 8 = \log \sqrt{48} = 2 + \log \sqrt{3}$.

$\xi_2 = (x_{1|1/4}, x_{2|1/4}, x_{3|1/12}, x_{4|1/12}, x_{5|1/12}, x_{6|1/12}, x_{7|1/12}, x_{8|1/12})$;

$H(\xi_2) = 2 \cdot \frac{1}{4} \log 4 + 6 \cdot \frac{1}{12} \log 12 = \log \sqrt{48} = 2 + \log \sqrt{3}$.

Энтропия — это матожидание количества информации, произведенного при генерации одного символа.

★ Энтропия источника M с множеством состояний S есть $H(M) = \sum_{s \in S} P(s) H(\xi_s)$, где $P(s)$ — вероятность нахождения источника в состоянии s в момент времени t, усредненная по t.

- вероятности $P(s)$ зависят от кубиков η, а не от кубиков ξ

Энтропия источника

Наша цель — найти подходящий к данным источник с маленькой энтропией. Мы пока умеем считать только энтропию случайной величины, т.е. марковского источника порядка 0, например:

$\xi_1 = (x_{1|1/6}, x_{2|1/6}, x_{3|1/6}, x_{4|1/8}, x_{5|1/8}, x_{6|1/8}, x_{7|1/8})$;

$H(\xi_1) = 3 \cdot \frac{1}{6} \log 6 + 4 \cdot \frac{1}{8} \log 8 = \log \sqrt{48} = 2 + \log \sqrt{3}$.

$\xi_2 = (x_{1|1/4}, x_{2|1/4}, x_{3|1/12}, x_{4|1/12}, x_{5|1/12}, x_{6|1/12}, x_{7|1/12}, x_{8|1/12})$;

$H(\xi_2) = 2 \cdot \frac{1}{4} \log 4 + 6 \cdot \frac{1}{12} \log 12 = \log \sqrt{48} = 2 + \log \sqrt{3}$.

Энтропия — это матожидание количества информации, произведенного при генерации одного символа.

★ Энтропия источника M с множеством состояний S есть $H(M) = \sum_{s \in S} P(s) H(\xi_s)$, где $P(s)$ — вероятность нахождения источника в состоянии s в момент времени t, усредненная по t.

- вероятности $P(s)$ зависят от кубиков η, а не от кубиков ξ

Откуда взять $P(s)$?

“Эргодическая” теорема Маркова

Пусть M — дискретная марковская цепь с m состояниями, к тому же сильно связная (т.е. для любых состояний s_1, s_2 процесс, находящийся в s_1, с ненулевой вероятностью в будущем посетит s_2). Тогда $(P(s_1), \ldots, P(s_m))$ — единственный левый собственный вектор матрицы M, принадлежащий собственному значению 1.

Пояснения к теореме Маркова

1. Теорема Перрона-Фробениуса говорит, что у неотрицательной матрицы есть “главное” собственное число: оно положительно и по модулю не меньше остальных собственных чисел; ему соответствует неотрицательный собственный вектор, единственный в случае “сильно связной” матрицы

Пояснения к теореме Маркова

1. Теорема Перрона-Фробениуса говорит, что у неотрицательной матрицы есть “главное” собственное число: оно положительно и по модулю не меньше остальных собственных чисел; ему соответствует неотрицательный собственный вектор, единственный в случае “сильно связной” матрицы
2. У стохастической матрицы M есть собственное число 1:
   - В матрице $M - E$ сумма элементов любой строки равна 0

   $\Rightarrow$ сумма столбцов $M - E$ равна нулевому вектору, т.е. столбцы линейно зависимы

   $\Rightarrow$ система $(M - 1 \cdot E)\vec{x} = \vec{0}$ имеет ненулевое решение

   $\Rightarrow$ 1 — собственное число M по определению

Пояснения к теореме Маркова

1. Теорема Перрона-Фробениуса говорит, что у неотрицательной матрицы есть "главное" собственное число: оно положительно и по модулю не меньше остальных собственных чисел; ему соответствует неотрицательный собственный вектор, единственный в случае "сильно связной" матрицы
2. У стохастической матрицы M есть собственное число 1:
   - В матрице $M - E$ сумма элементов любой строки равна 0

   $\Rightarrow$ сумма столбцов $M - E$ равна нулевому вектору, т.е. столбцы линейно зависимы

   $\Rightarrow$ система $(M - 1 \cdot E)\vec{x} = \vec{0}$ имеет ненулевое решение

   $\Rightarrow$ 1 — собственное число M по определению
3. Главное собственное число стохастической матрицы M равно 1:
   - При умножении вектора-строки на стохастическую матрицу сумма координат вектора не меняется:
     $(x_1, \ldots, x_m)M = (x_1 M[1,1] + \cdots + x_m M[m,1], \ldots, x_1 M[1,m] + \cdots + x_m M[m,m])$,
     при сложении координат каждый x_i умножается на сумму
     $M[i,1] + \cdots + M[i,m] = 1$
   - при умножении собственного вектора на M каждая координата умножается на собственное число

   $\Rightarrow$ каждый собственный вектор либо принадлежит собственному числу 1, либо имеет нулевую сумму координат

   $\Rightarrow$ по теореме Перрона-Фробениуса у главного собственного числа есть собственный вектор с ненулевой суммой координат

   $\Rightarrow$ 1 — единственный кандидат на главное собственное число M

Пояснения к теореме Маркова

1. Теорема Перрона-Фробениуса говорит, что у неотрицательной матрицы есть "главное" собственное число: оно положительно и по модулю не меньше остальных собственных чисел; ему соответствует неотрицательный собственный вектор, единственный в случае "сильно связной" матрицы
2. У стохастической матрицы M есть собственное число 1:
   - В матрице $M - E$ сумма элементов любой строки равна 0

   $\Rightarrow$ сумма столбцов $M - E$ равна нулевому вектору, т.е. столбцы линейно зависимы

   $\Rightarrow$ система $(M - 1 \cdot E)\vec{x} = \vec{0}$ имеет ненулевое решение

   $\Rightarrow$ 1 — собственное число M по определению
3. Главное собственное число стохастической матрицы M равно 1:
   - При умножении вектора-строки на стохастическую матрицу сумма координат вектора не меняется:
     $(x_1, \ldots, x_m)M = (x_1 M[1,1] + \cdots + x_m M[m,1], \ldots, x_1 M[1,m] + \cdots + x_m M[m,m])$,
     при сложении координат каждый x_i умножается на сумму
     $M[i,1] + \cdots + M[i,m] = 1$
   - при умножении собственного вектора на M каждая координата умножается на собственное число

   $\Rightarrow$ каждый собственный вектор либо принадлежит собственному числу 1, либо имеет нулевую сумму координат

   $\Rightarrow$ по теореме Перрона-Фробениуса у главного собственного числа есть собственный вектор с ненулевой суммой координат

   $\Rightarrow$ 1 — единственный кандидат на главное собственное число M

   ★ Вероятности $P(s_1), \ldots, P(s_m)$ можно найти, решив однородную систему уравнений $\vec{x}(M - E) = 0$

Сжатие с антисловарем

Сжатие с антисловарем

Сжатие с антисловарем — "эзотерический" метод, не получивший практического применения, но удобный для понимания, построения марковского источника и вычисления энтропии оного

Пусть $T \in \Sigma^n$ — текст.

- Строка $u \in \Sigma^*$ называется запрещенной в T, если она не входит в T как подстрока
- Множество строк называется антифакториальным, если для любой пары строк u, v в нем u — не подстрока v
- Антисловарь для T — произвольное антифакториальное множество A запрещенных в T подстрок

Пример: $T = 1011010110$, $A = \{00, 111, 01010\}$

Сжатие с антисловарем

Сжатие с антисловарем — "эзотерический" метод, не получивший практического применения, но удобный для понимания, построения марковского источника и вычисления энтропии оного

Пусть $T \in \Sigma^n$ — текст.

- Строка $u \in \Sigma^*$ называется запрещенной в T, если она не входит в T как подстрока
- Множество строк называется антифакториальным, если для любой пары строк u, v в нем u — не подстрока v
- Антисловарь для T — произвольное антифакториальное множество A запрещенных в T подстрок

Пример: $T = 1011010110$, $A = \{00, 111, 01010\}$

- В алгоритме сжатия с антисловарем текст рассматривается как последовательность бит: $\Sigma = \{0, 1\}$
- Некоторые биты можно предсказать, зная предыдущий текст:
  - если предыдущий текст заканчивался подстрокой u и в антисловаре есть $u0$, то очередной бит точно равен 1
- Сжатие состоит в вычеркивании из текста всех предсказуемых битов

Алгоритмы сжатия и разжатия по антисловарю

Пусть антисловарь A фиксирован
Текст T — любой, для которого A — антисловарь

Алгоритм сжатия (вход: текст T; выход: сжатый текст C, число $n = |T|$)

- $C \leftarrow \lambda$; $i, j \leftarrow 1$; пока (не конец T) повторять
  - $a \leftarrow T[i]$
  - если $T[1..i{-}1]\bar{a}$ не заканчивается строкой из A, то $C[j] \leftarrow a$; $j \leftarrow j{+}1$
  - $i \leftarrow i + 1$
- вернуть C, i

Алгоритмы сжатия и разжатия по антисловарю

Пусть антисловарь A фиксирован
Текст T — любой, для которого A — антисловарь

Алгоритм сжатия (вход: текст T; выход: сжатый текст C, число $n = |T|$)

- $C \leftarrow \lambda$; $i, j \leftarrow 1$; пока (не конец T) повторять
  - $a \leftarrow T[i]$
  - если $T[1..i-1]\bar{a}$ не заканчивается строкой из A, то $C[j] \leftarrow a$; $j \leftarrow j+1$
  - $i \leftarrow i + 1$
- вернуть C, i

Пример: $A = \{00, 111, 01010\}$
T 1011010110
C 10 1 0

Алгоритмы сжатия и разжатия по антисловарю

Пусть антисловарь A фиксирован
Текст T — любой, для которого A — антисловарь

Алгоритм сжатия (вход: текст T; выход: сжатый текст C, число $n = |T|$)

- $C \leftarrow \lambda$; $i, j \leftarrow 1$; пока (не конец T) повторять
  - $a \leftarrow T[i]$
  - если $T[1..i-1]\bar{a}$ не заканчивается строкой из A, то $C[j] \leftarrow a$; $j \leftarrow j+1$
  - $i \leftarrow i + 1$
- вернуть C, i

Пример: $A = \{00, 111, 01010\}$

T	1011010110	запрещенные слова	00 111 01010
C	10 1 0	предсказанные биты	$3, 6, 8$ $5, 10$ 9

Алгоритмы сжатия и разжатия по антисловарю

Пусть антисловарь A фиксирован
Текст T — любой, для которого A — антисловарь

Алгоритм сжатия (вход: текст T; выход: сжатый текст C, число $n = |T|$)

- $C \leftarrow \lambda$; $i, j \leftarrow 1$; пока (не конец T) повторять
  - $a \leftarrow T[i]$
  - если $T[1..i{-}1]\bar{a}$ не заканчивается строкой из A, то $C[j] \leftarrow a$; $j \leftarrow j{+}1$
  - $i \leftarrow i + 1$
- вернуть C, i

Пример: $A = \{00, 111, 01010\}$

T	1011010110	запрещенные слова	00 111 01010
C	10 1 0	предсказанные биты	$3, 6, 8$ $5, 10$ 9

Алгоритм разжатия (вход: сжатый текст C, число $n = |T|$; выход: текст T)

- $T \leftarrow \lambda$; $j \leftarrow 1$; $i \leftarrow |T|$; пока $i < n$ повторять
  - если $T[1..i]a$ для некоторого бита a заканчивается строкой из A, то $T[i{+}1] \leftarrow \bar{a}$
  - иначе $T[i{+}1] \leftarrow C[j]$; $j \leftarrow j{+}1$
  - $i \leftarrow i{+}1$
- вернуть T

Антисловарь и конечные автоматы

- Как эффективно (за время $O(|T|)$) реализовать сжатие и разжатие?
  - Построим конечный автомат-преобразователь (1 шаг на бит входа)
- Что за источник у текстов с антисловарем A и какова его энтропия?
  - “Допилим” автомат до марковского источника

Антисловарь и конечные автоматы

- Как эффективно (за время $O(|T|)$) реализовать сжатие и разжатие?
  - Построим конечный автомат-преобразователь (1 шаг на бит входа)
- Что за источник у текстов с антисловарем A и какова его энтропия?
  - “Допилим” автомат до марковского источника

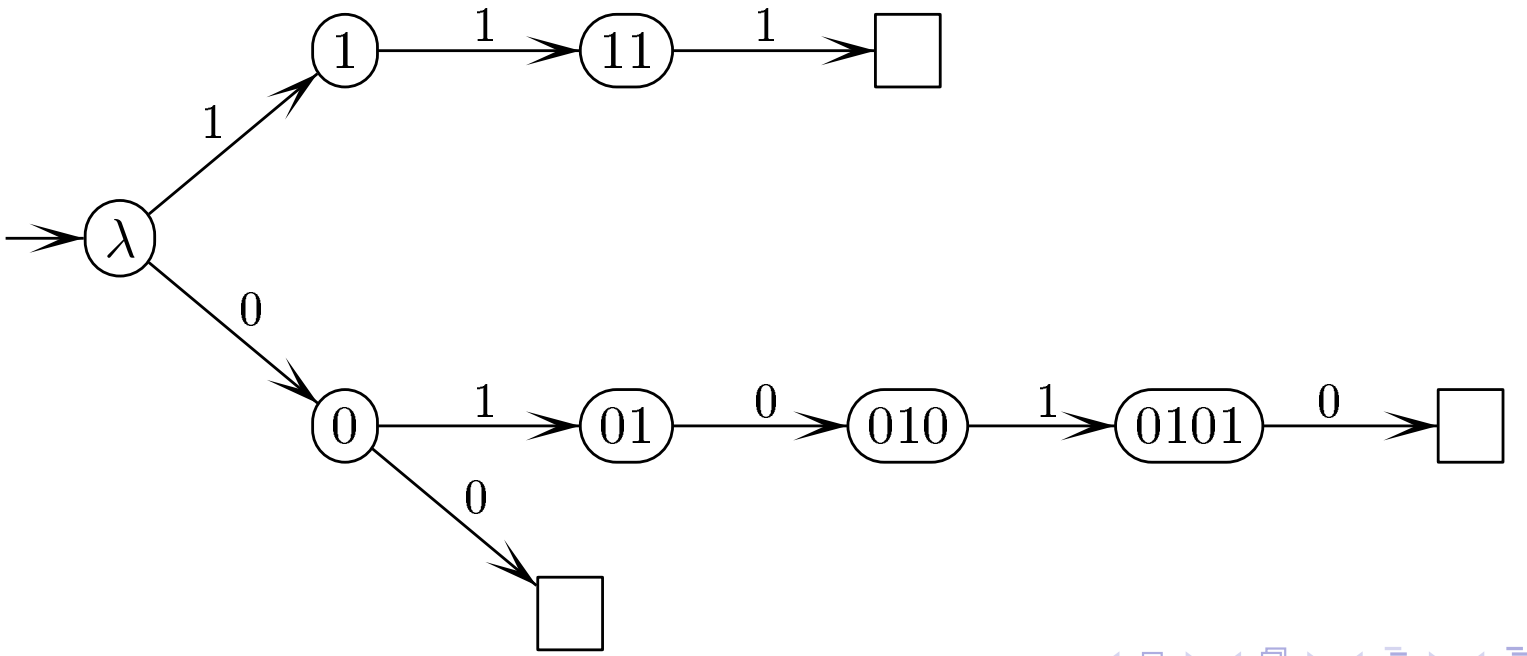
- Автомат должен распознавать все строки, не имеющие подстрок из A
- Такой автомат-”спамофильтр” можно построить по A — это модификация автомата Ахо-Корасик

Антисловарь и конечные автоматы

- Как эффективно (за время $O(|T|)$) реализовать сжатие и разжатие?
  - Построим конечный автомат-преобразователь (1 шаг на бит входа)
- Что за источник у текстов с антисловарем A и какова его энтропия?
  - "Допилим" автомат до марковского источника

- Автомат должен распознавать все строки, не имеющие подстрок из A
- Такой автомат-"спамофильтр" можно построить по A — это модификация автомата Ахо-Корасик

Строим древовидный автомат (бор), распознающий A:

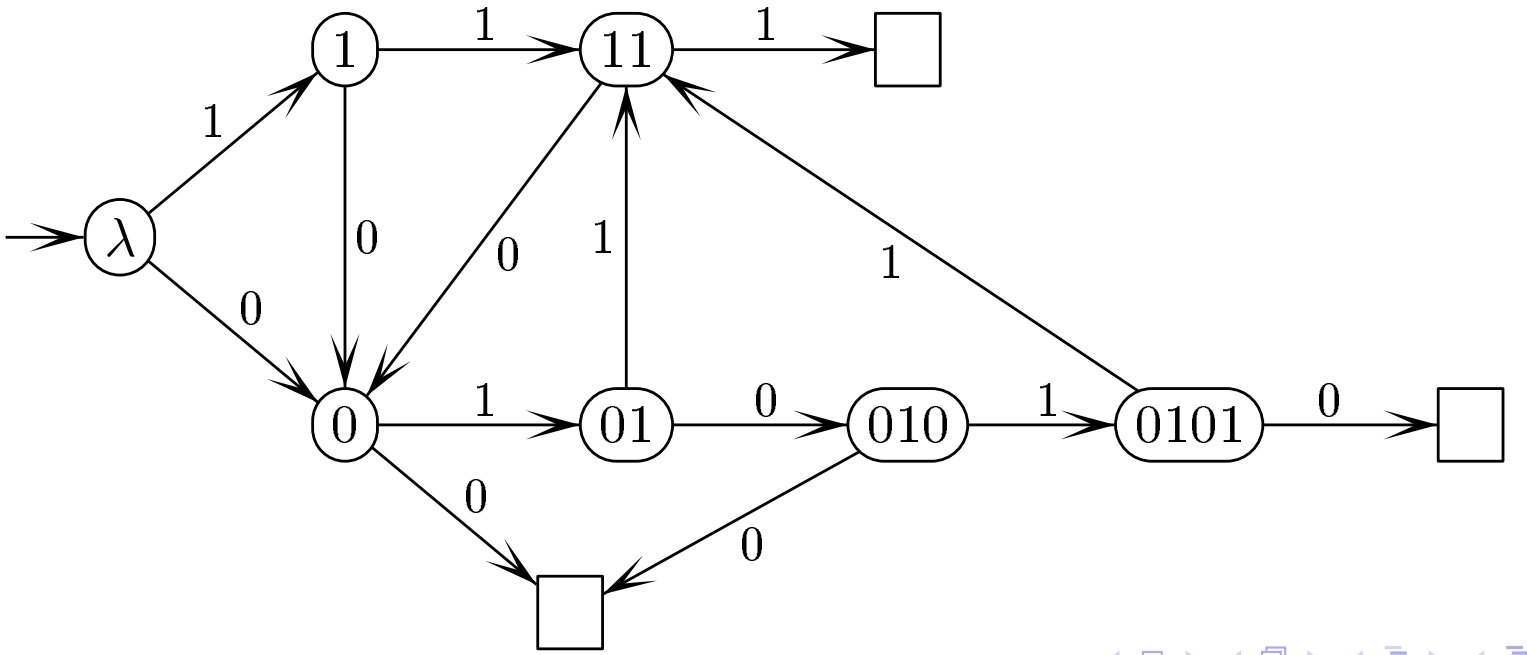


Антисловарь и конечные автоматы

- Как эффективно (за время $O(|T|)$) реализовать сжатие и разжатие?
  - Построим конечный автомат-преобразователь (1 шаг на бит входа)
- Что за источник у текстов с антисловарем A и какова его энтропия?
  - "Допилим" автомат до марковского источника

- Автомат должен распознавать все строки, не имеющие подстрок из A
- Такой автомат-"спамофильтр" можно построить по A — это модификация автомата Ахо-Корасик

Добавляем недостающие ребра ($u \xrightarrow{a} v$ если v — длиннейший суффикс ua в боре):

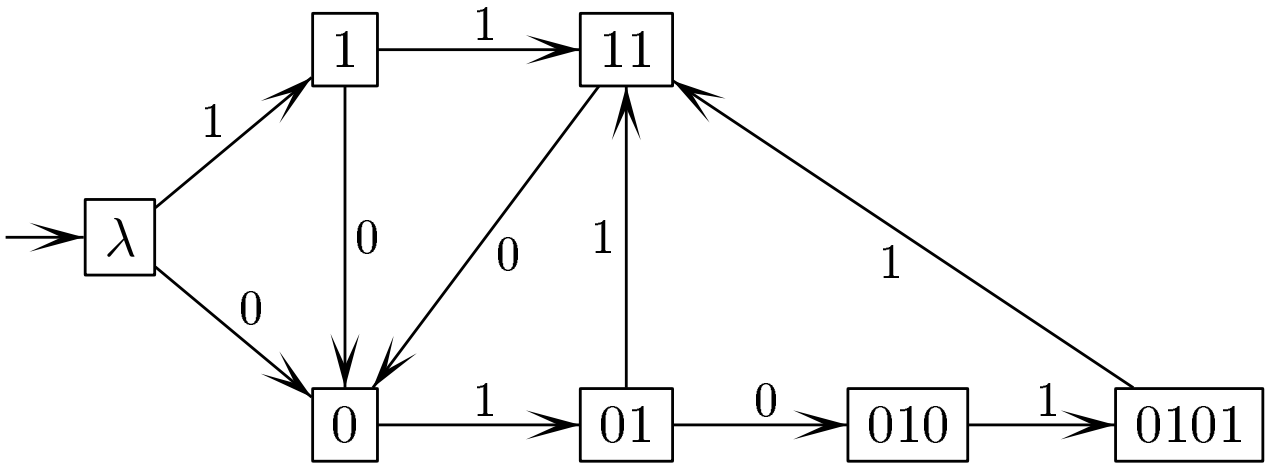


Антисловарь и конечные автоматы

- Как эффективно (за время $O(|T|)$) реализовать сжатие и разжатие?
  - Построим конечный автомат-преобразователь (1 шаг на бит входа)
- Что за источник у текстов с антисловарем A и какова его энтропия?
  - "Допилим" автомат до марковского источника

- Автомат должен распознавать все строки, не имеющие подстрок из A
- Такой автомат-"спамофильтр" можно построить по A — это модификация автомата Ахо-Корасик

Удаляем стоки, получая неполный ДКА, распознающий $\Sigma^* \setminus \Sigma^* A \Sigma^*$:

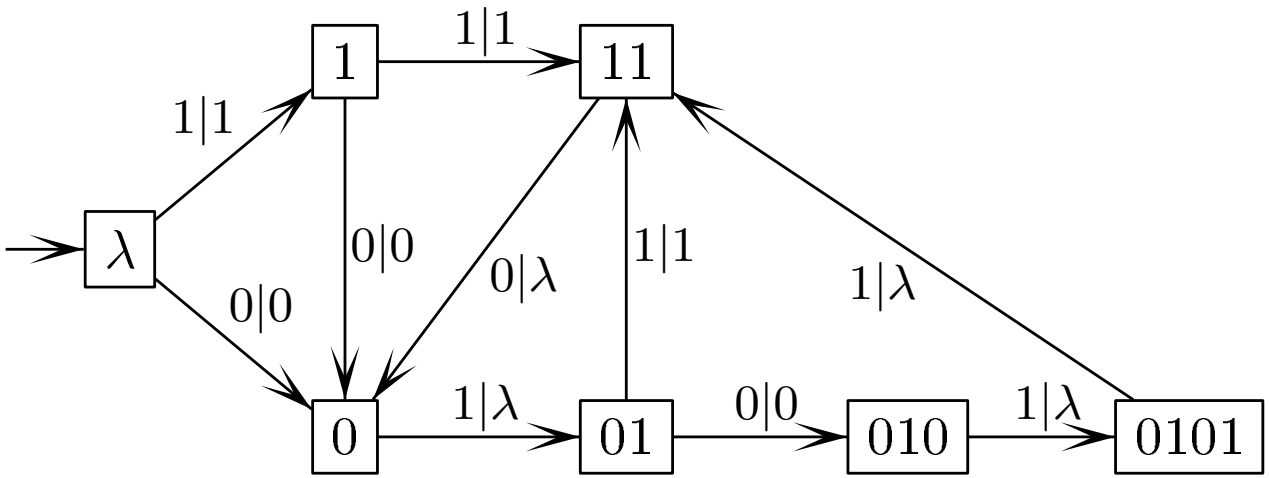


Антисловарь и конечные автоматы

- Как эффективно (за время $O(|T|)$) реализовать сжатие и разжатие?
  - Построим конечный автомат-преобразователь (1 шаг на бит входа)
- Что за источник у текстов с антисловарем A и какова его энтропия?
  - "Допилим" автомат до марковского источника

- Автомат должен распознавать все строки, не имеющие подстрок из A
- Такой автомат-"спамофильтр" можно построить по A — это модификация автомата Ахо-Корасик

Конечный преобразователь для сжатия:

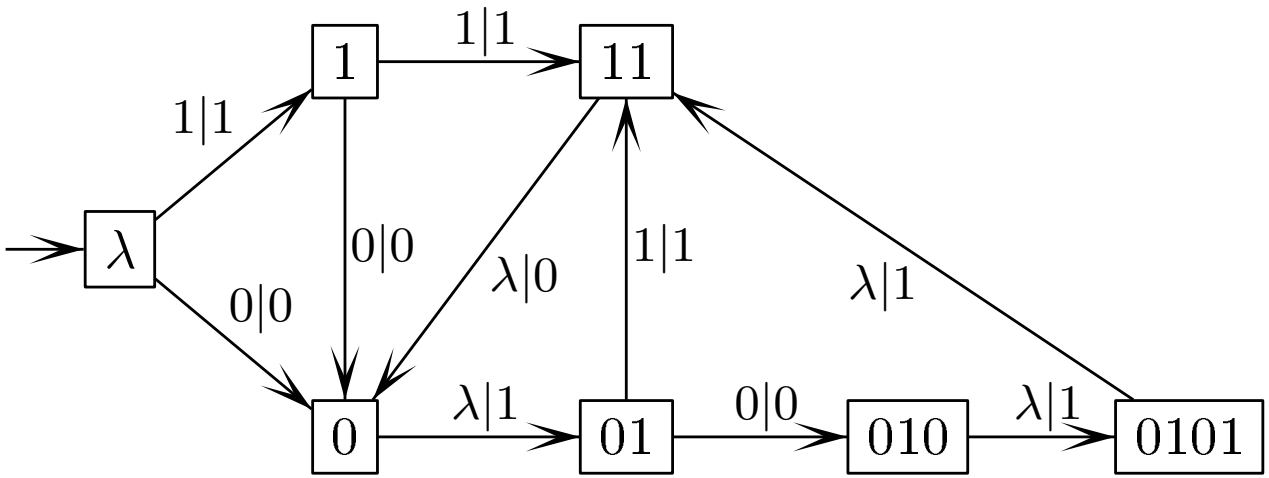


Антисловарь и конечные автоматы

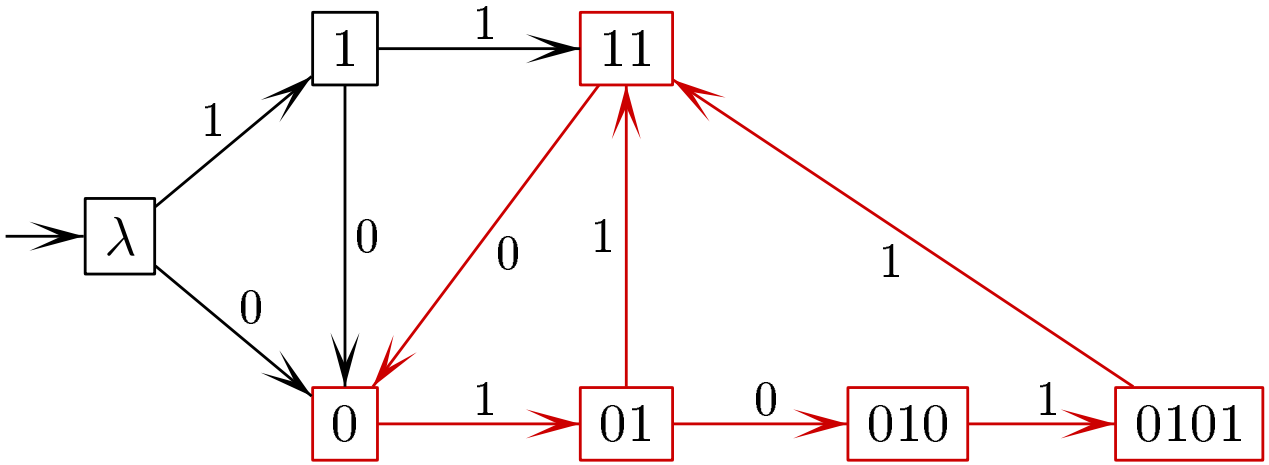
- Как эффективно (за время $O(|T|)$) реализовать сжатие и разжатие?
  - Построим конечный автомат-преобразователь (1 шаг на бит входа)
- Что за источник у текстов с антисловарем A и какова его энтропия?
  - "Допилим" автомат до марковского источника

- Автомат должен распознавать все строки, не имеющие подстрок из A
- Такой автомат-"спамофильтр" можно построить по A — это модификация автомата Ахо-Корасик

Конечный преобразователь для разжатия:

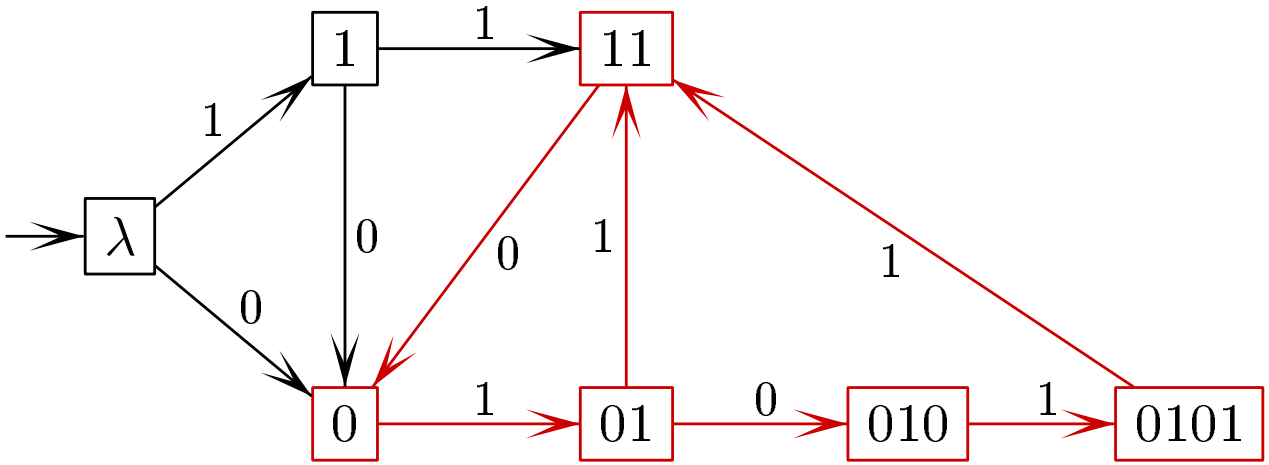


Энтропия источника для сжатия с антисловарем



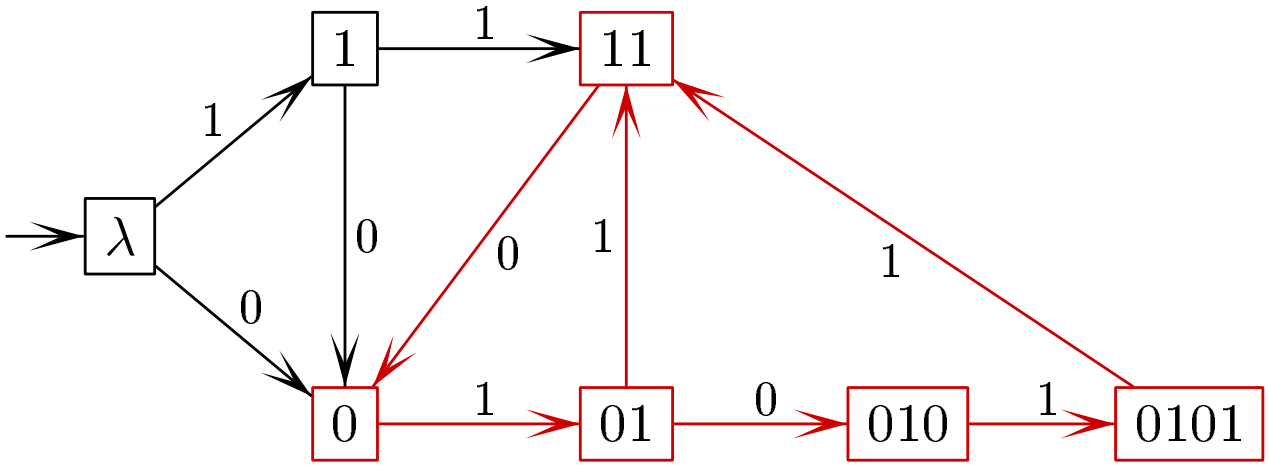
- Построенный автомат — марковский источник S смешанного порядка:
  - состояния являются левыми контекстами текущего символа
  - если из s выходит одна стрелка и a — ее метка, то $\xi_s = (a_{|1})$ (константа)
  - если из s выходят две стрелки, то полагаем $\xi_s = (0_{|1/2}, 1_{|1/2})$ (монетка)

Энтропия источника для сжатия с антисловарем



- Построенный автомат — марковский источник S смешанного порядка:
  - состояния являются левыми контекстами текущего символа
  - если из s выходит одна стрелка и a — ее метка, то $\xi_s = (a_{|1})$ (константа)
  - если из s выходят две стрелки, то полагаем $\xi_s = (0_{|1/2}, 1_{|1/2})$ (монетка)
- $H(S) = \sum_s P(S)$, где суммирование ведется по всем состояниям с монетками
- В нашем примере можно исключить состояния λ и 1 с $P = 0$ (в них нельзя попасть начиная с $t = 2$); остальные состояния (красный подграф) образуют сильно связную цепь с теми же вероятностями, что и исходная

Энтропия источника для сжатия с антисловарем



- Построенный автомат — марковский источник S смешанного порядка:
  - состояния являются левыми контекстами текущего символа
  - если из s выходит одна стрелка и a — ее метка, то $\xi_s = (a_{|1})$ (константа)
  - если из s выходят две стрелки, то полагаем $\xi_s = (0_{|1/2}, 1_{|1/2})$ (монетка)
- $H(S) = \sum_s P(S)$, где суммирование ведется по всем состояниям с монетками
- В нашем примере можно исключить состояния λ и 1 с $P = 0$ (в них нельзя попасть начиная с $t = 2$); остальные состояния (красный подграф) образуют сильно связную цепь с теми же вероятностями, что и исходная

Осталось найти вероятности для цепи с матрицей $M = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{matrix} 11 \\ 0 \\ 01 \\ 010 \\ 0101 \end{matrix}$

Решаем систему $\vec{x}(M-E)=\vec{0}$, т.е. однородную систему с матрицей $(M-E)^{\perp}=$

$$\begin{pmatrix} -1 & 0 & \frac{1}{2} & 0 & 1 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} \sim \begin{pmatrix} -1 & 0 & \frac{1}{2} & 0 & 1 \\ 0 & -1 & \frac{1}{2} & 0 & 1 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} \sim \begin{pmatrix} -1 & 0 & \frac{1}{2} & 0 & 1 \\ 0 & -1 & \frac{1}{2} & 0 & 1 \\ 0 & 0 & -\frac{1}{2} & 0 & 1 \\ 0 & 0 & \frac{1}{2} & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} \sim$$

$$\begin{pmatrix} -1 & 0 & \frac{1}{2} & 0 & 1 \\ 0 & -1 & \frac{1}{2} & 0 & 1 \\ 0 & 0 & -\frac{1}{2} & 0 & 1 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} \sim \begin{pmatrix} -1 & 0 & 0 & 0 & 2 \\ 0 & -1 & 0 & 0 & 2 \\ 0 & 0 & -1 & 0 & 2 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Вычисление вероятностей

Решаем систему $\vec{x}(M-E)=\vec{0}$, т.е. однородную систему с матрицей $(M-E)^{\perp}=$

$$\begin{pmatrix} -1 & 0 & \frac{1}{2} & 0 & 1 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} \sim \begin{pmatrix} -1 & 0 & \frac{1}{2} & 0 & 1 \\ 0 & -1 & \frac{1}{2} & 0 & 1 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} \sim \begin{pmatrix} -1 & 0 & \frac{1}{2} & 0 & 1 \\ 0 & -1 & \frac{1}{2} & 0 & 1 \\ 0 & 0 & -\frac{1}{2} & 0 & 1 \\ 0 & 0 & \frac{1}{2} & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} \sim$$

$$\begin{pmatrix} -1 & 0 & \frac{1}{2} & 0 & 1 \\ 0 & -1 & \frac{1}{2} & 0 & 1 \\ 0 & 0 & -\frac{1}{2} & 0 & 1 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} \sim \begin{pmatrix} -1 & 0 & 0 & 0 & 2 \\ 0 & -1 & 0 & 0 & 2 \\ 0 & 0 & -1 & 0 & 2 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

С дополнительным условием $x_1+x_2+x_3+x_4+x_5=1$ получаем $x_1=x_2=x_3=1/4$, $x_4=x_5=1/8$, $H(S)=x_3=1/4$, $|C|/|T|\to 1/4$

Вычисление вероятностей

Решаем систему $\vec{x}(M-E)=\vec{0}$, т.е. однородную систему с матрицей $(M-E)^{\perp}=$

$$\begin{pmatrix} -1 & 0 & \frac{1}{2} & 0 & 1 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} \sim \begin{pmatrix} -1 & 0 & \frac{1}{2} & 0 & 1 \\ 0 & -1 & \frac{1}{2} & 0 & 1 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} \sim \begin{pmatrix} -1 & 0 & \frac{1}{2} & 0 & 1 \\ 0 & -1 & \frac{1}{2} & 0 & 1 \\ 0 & 0 & -\frac{1}{2} & 0 & 1 \\ 0 & 0 & \frac{1}{2} & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} \sim$$

$$\begin{pmatrix} -1 & 0 & \frac{1}{2} & 0 & 1 \\ 0 & -1 & \frac{1}{2} & 0 & 1 \\ 0 & 0 & -\frac{1}{2} & 0 & 1 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} \sim \begin{pmatrix} -1 & 0 & 0 & 0 & 2 \\ 0 & -1 & 0 & 0 & 2 \\ 0 & 0 & -1 & 0 & 2 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

С дополнительным условием $x_1+x_2+x_3+x_4+x_5=1$ получаем
$x_1=x_2=x_3=1/4$, $x_4=x_5=1/8$, $H(S)=x_3=1/4$, $|C|/|T|\to 1/4$

★ Почему сжатие с антисловарем — непрактический метод?

- Плохо отражает закономерности в реальных текстах, и поэтому проигрывает другим методам в степени сжатия
- Нужно строить антисловарь по каждому тексту и хранить/передавать со сжатым текстом; для хорошего сжатия антисловарь должен быть большим

Дальнейшие планы

Часть первая: учимся кодировать

- Префиксное кодирование
- Арифметическое кодирование
- Кодирование чисел и числовых последовательностей

Часть вторая: учимся моделировать

- Статистическое кодирование; метод PPM
- Словарное кодирование; методы Лемпеля-Зива (LZ)
- Комбинаторный трюк: методы, основанные на преобразовании Барроуза-Уилера (BWT)