§3. Поиск в тексте вхождения слова или множества слов. Алгоритм Кнута-Мориса-Прата. Автомат Ахо-Корасик.

Пусть $t \in \Sigma^*$, в дальнейшем называется текстом.
Задача: для данного слова *w* найти его вхождение в *t* (или все вхождения).

Замечание: такую задачу нужно решать для фильтрации сообщений с признаками спама/нецензурного контента. Или для выборки научных статей по интересующей теме.

Алгоритм грубой силы:
Пусть i – первая позиция в тексте t.
Сравниваем буквы в t с позиции i и в w с позиции 1.

t_i	t_{i+1}	$\cdots$		
w_1	w_2	$\cdots$		

Если совпали буквы до последней в w, то i – выход.

Если $t_{i+k} \neq w_k$, то увеличиваем i, и повторяем.

Сложность такого алгоритма имеет порядок $O(|t| \cdot |w|)$.

В алгоритме Кнута-Мориса-Прата существенно используется префикс-функция.

Опр. Префикс-функцией для слова $w = \alpha_1\alpha_2...\alpha_m$ называется $\pi : \{1,2,..,m\} \to \{0,1,2,..,m-1\}$, такая, что $\pi(k)$ равна длине наибольшего собственного префикса подслова $\alpha_1\alpha_2...\alpha_k$, являющегося суффиксом этого подслова.

Пример. $w = abab$.

k	$\pi(k)$	префикс = суффикс
1	0	
2	0	
3	1	a
4	2	ab

Алгоритм Кнута-Мориса-Прата.
Вход: t, π для w.
1) Присвоить значение $j = 0$.
2) Цикл по i с 1 до n (длина t).

Повторять, пока не выйдем из условия $t_i \neq \alpha_{j+1}$ и $j > 0$,

присвоение $j = \pi(j)$.

Если $t_i = \alpha_{j+1}$, то увеличить j на 1.

Если $j = m$ (длина w), то присвоить $j = \pi(j)$. Подстрока найдена. Выход – $(i - m)$.

Запись алгоритма в виде программы на языке высокого уровня (без соблюдения синтаксиса):

j **:=** 0
for i =1 **to** n
  while ($t_i \neq \alpha_{j+1}$ и $j > 0$) **do**
    j**:=** $\pi(j)$
  if $t_i = \alpha_{j+1}$ **then** j **:=** $j + 1$
  if $j = m$ **then**
    $j = \pi(j)$
    return $(i - m)$
% конец

Лемма (без док-ва). При условии использования уже вычисленной функции π алгоритм Кнута-Мориса-Прата работает за время $\Theta(n)$.

Замечание: для вычисления функции π требуется найти вхождения префиксов слова w в себя.

Реализация алгоритма Кнута-Мориса-Прата при помощи автомата КМР.

Пусть $Q = (0, 1, \ldots, m\}$.

0 – начальное состояние.

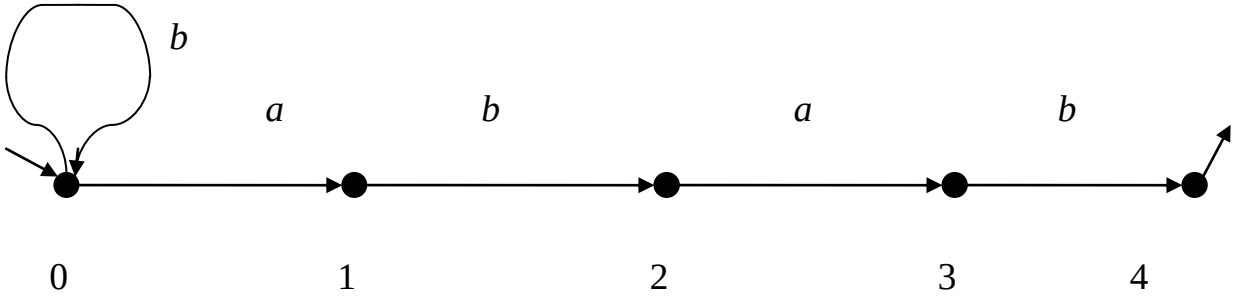
$$\varphi(0,x) = \begin{cases} 0, & \text{если } \alpha_1 \neq x \\ 1, & \text{если } \alpha_1 = x \end{cases}.$$

Для всех $j = 1, ..., m - 1$

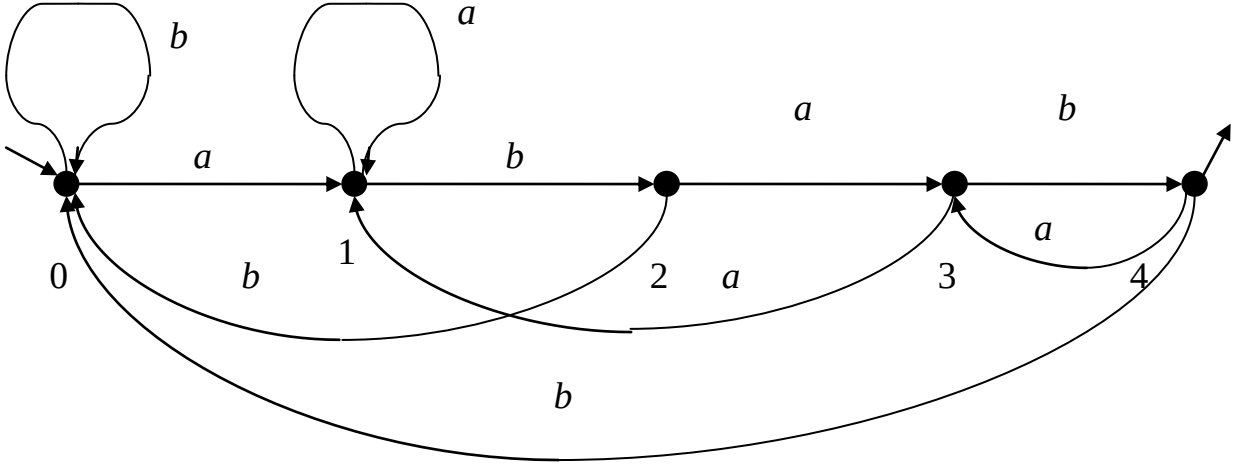
$$\varphi(j,x) = \begin{cases} \varphi(\pi(j),x), & \text{если } \alpha_j \neq x \\ j+1, & \text{если } \alpha_j = x \end{cases}.$$

$\varphi(m,x) = \varphi(\pi(m),x)$. $Q_F = \{m\}$.

Пример. $w = abab$.



Добавим переходы $\varphi(1,a)=\varphi(\pi(1),a)=\varphi(0,a)=1$,
$\varphi(2,b)=\varphi(\pi(2),b)=\varphi(0,b)=0$, $\varphi(3,a)=\varphi(\pi(3),a)=\varphi(1,a)=1$,
$\varphi(4,a)=\varphi(\pi(4),a)=\varphi(2,a)=3$, $\varphi(4,b)=\varphi(\pi(4),b)=\varphi(2,b)=0$.



Теорема (без док-ва). Автомат КМР допускает язык $L = \Sigma^* w$, состоящий из всех слов, заканчивающихся на слово w.

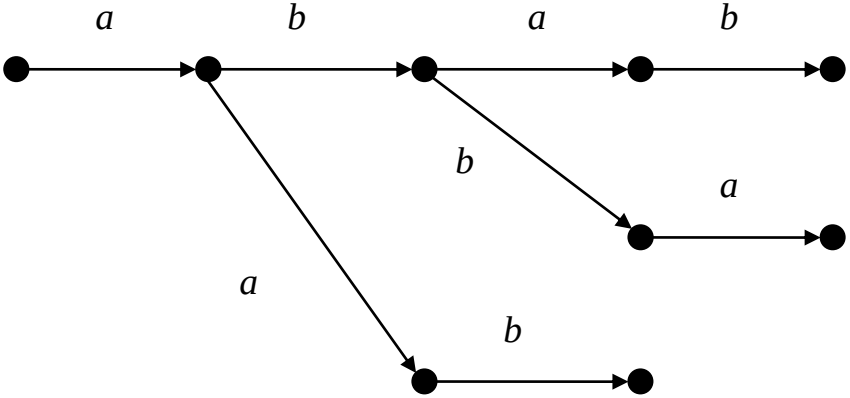
Пусть $W = \{w_1, w_2, ..., w_p\}$ конечное множество слов, «словарь».
Задача: для данного словаря найти все вхождения слов из словаря в *t*.

Алгоритм Ахо-Корасик использует обобщение префикс-функции и соответствующий автомат.

Опр. Префиксным деревом для множества слов $W = \{w_1, w_2, ..., w_p\}$ называется дерево, ребра в котором помечены буквами слов из *W*, так, что путь от корня к вершине дерева соответствует префиксу некоторого слова из *W*.

Замечание: в Википедии используется термин «бор» для префиксного дерева.

Пример. $W = \{abab, abba, aab\}$.



Автомат, допускающий язык $L = \Sigma^* W$, строится на основе префиксного дерева. Заключительными состояниями будут вершины дерева, соответствующие словам из W.

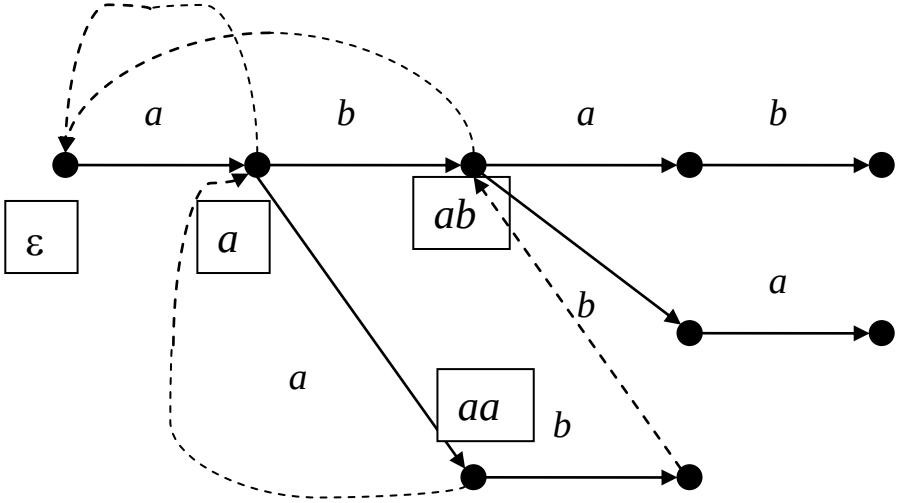
К нему добавляются переходы, использующие обобщенную префикс-функцию.

Рекуррентное определение обобщенной префикс-функции.

Пусть π: $\Sigma^* \to \Sigma^*$ (не обязательно всюду определенная).

1. Для каждого слова u длины 1, у которого есть путь в префиксном дереве, $\pi(u) = \varepsilon$.
2. Если для слова $u\alpha$ (где $\alpha \in \Sigma$) есть путь в префиксном дереве, и $\pi(u)$ уже определен, то $\pi(u\alpha) = \pi^{\ell}(u)\alpha$, где ℓ – наименьшая степень, такая, что $\pi^{\ell}(u)\alpha$ имеет путь в префиксном дереве, либо ε (если $\pi^{\ell}(u)\alpha$ не определен).

Пример. Пунктирной линией будем обозначать значения обобщенной префиксной функции.



$\pi(a) = \varepsilon$. $\pi(ab) = \varepsilon$. $\pi(aa) = a$. $\pi(aab) = ab$. И так далее.

u	$\pi(u)$ – первый шаг	$\pi(u)$ – второй шаг	$\pi(u)$ – третий шаг	$\pi(u)$ – четвертый шаг
a	ε	ε	ε	ε
ab		ε	ε	ε
aa		a	a	a
aba			a	a
abb			ε	ε
aab			ab	ab
$abab$				ab
$abba$				a

Автомат, допускающий язык $L = \Sigma^* W$, имеет множество состояний Q, равное количеству вершин в префиксном дереве для словаря W.

Начальное состояние q_0 – корень дерева.

Заключительные состояния – вершины дерева, соответствующие словам из W.

Каждое помеченное буквой x ребро префиксного дерева будет результатом функции переходов $\varphi(q, x)$.

Если в префиксном дереве нет перехода $\varphi(q_0, x)$, то назначаем $\varphi(q_0, x) = q_0$.

Если в префиксном дереве нет перехода $\varphi(q, x)$, то назначаем $\varphi(q, x) = \varphi(\pi^{\ell}(q), x)$ или $\varphi(q, x) = q_0$.

Пример.

	a	*b*	закл.
ε	*a*	ε	0
a	*aa*	*ab*	0
ab	*aba*	*abb*	0
aa	*aa*	*aab*	0
aba	*aa*	*abab*	0
abb	*abba*	ε	1
aab	*aba*	*abb*	0
abab	*aba*	*abb*	1
abba	*aa*	*ab*	1