Секция «Математика и механика»

Нижние оценки временной и объёмной сложности задачи поиска подстроки

Перпер Евгений Михайлович
Студент
Московский государственный университет имени М.В. Ломоносова,
Механико-математический факультет, Москва, Россия
E-mail: e_m_perper@mail.ru

Задача, рассматриваемая в данной работе, состоит в следующем. Пусть база данных состоит из некоторого множества слов, запросом к базе данных тоже является слово, и мы хотим по запросу получить все слова из базы данных, в которых этот запрос содержится в качестве подслова, и не получить никаких других слов. Данная задача может возникнуть, например, при поиске в каком-нибудь словаре или тексте слов, имеющих тот же корень, что и данное слово. Мы будем рассматривать алгоритмы, решающие эту задачу быстро, за время, пропорциональное длине запроса. Некоторые из этих алгоритмов уже рассматривались в [1].

Для каждого алгоритма база данных состоит из слов, которые в ней имелись изначально (назовём множество этих слов словарём и обозначим его как V), а также из некоторых дополнительных элементов. Длину слова w обозначим через $l(w)$. Считаем, что все слова из V имеют одну и ту же длину, а длина каждого запроса x не превышает длину слова из V. Множество всех допустимых запросов обозначим через X.

Обработка запроса будет заключаться в переходе по ссылкам между элементами базы данных в результате вычисления некоторых функций. Мы будем рассматривать класс алгоритмов, использующих лишь функцию $f(x)$, тождественно равную 1, а также функции $g_i(x), i \in N$, выдающие в качестве результата i-ю букву запроса, либо специальный символ, если i превышает длину запроса. Считаем, что время вычисления $f(x)$ и последующего перехода по ссылке равно t, $0 < t < 1$, а время вычисления функции $g_i(x)$ и последующего перехода по ссылке равно 1. Общее время обработки запроса x алгоритмом u обозначим как $T(u,x)$, а объём памяти алгоритма, вычисляемый как количество ссылок между элементами соответствующей базы данных, обозначим как $Q(u)$. Обозначим класс алгоритмов, решающих задачу поиска подстроки при заданном V, притом использующих только функции из множества $F = \{f\} \cup \{g_i, i \in N\}$, через $U(V,F)$. Пусть $d(V,x)$ — количество слов в V, удовлетворяющих запросу x, а n – длина каждого слова из V. Рассмотрим функцию $R(V,F,x)$, равную n, если $l(x) = n$, и равную $l(x) + 1 + t(d(V,x) - 1)$, если $l(x) < n$.

Теорема 1. *Пусть в алфавите $k \geq 3$ различных букв. Тогда среди всех словарей из p слов, имеющих длину $n \geq 3$, где $p \leq (k-1)^{[n/3]}$, найдётся такой словарь V, что $\forall u \in U(V,F), \forall x \in X \quad d(V,x) \geq 1 \Rightarrow T(u,x) \geq R(V,F,x)$.*

Обозначим $U_V = \{u \in U(V,F) : \forall x \in X \; d(V,x) \geq 1 \Rightarrow T(u,x) \leq R(V,F,x)\}$. Пусть $Q(p,n) = \max_{V:|V|=p} \min_{u \in U_V} Q(u)$.

Теорема 2. *Пусть $k \geq 3$, $p \leq (k-1)^{[n/3]}$. Тогда $pn^2/9 \lesssim Q(p,n) \lesssim (k+1)pn^2/4$, $n \to \infty$, $p \to \infty$.*

Литература

1. Перпер Е.М. О сложности поиска подстроки// Интеллектуальные системы. 2012. Т. 16, вып. 1-4. С. 299-320.

Слова благодарности

Автор выражает благодарность профессору Э.Э.Гасанову за научное руководство.