

Алгоритмы и анализ трудоемкости обработки сжатых текстов

Лифшиц Юрий Михайлович
<http://logic.pdmi.ras.ru/~yura>

диссертация по специальности 05.13.17 —
"Теоретические основы информатики"

31 мая 2007 г.

1 / 11

Обработка сжатых текстов

Цели:

Для классических задач на строках определить, можно ли их решить за время, полиномиальное относительно размера архивов?

Если такой алгоритм найти не удастся, показать NP-трудность рассматриваемых задач

3 / 11

Тематика работы

- 1 Вычислительная сложность задач на строках в ситуации, когда на вход алгоритму строки попадают в **заархивированном** виде
- 2 **Разреженная периодичность** как система компактного описания строк на основе частично определенных слов

2 / 11

Актуальность

- Начало области — работа Амира, Бенсона и Фараха в 1994 году
- С тех пор — десятки работ, тема обработки сжатых текстов постоянно рассматривается на крупнейшей специализированной конференции: Combinatorial Pattern Matching
- Применения в верификации программ (Генест и Мушоль), уравнениях в словах (Пландовский), определении эквивалентности программ (Риттер и Ласота)

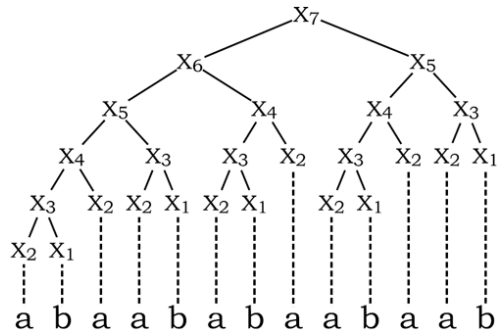
4 / 11

Прямолинейные программы

Большинство используемых методов архивирования (семейство Lempel-Ziv, все словарные методы) могут быть эффективно преобразованы в **прямолинейные программы**

Пример: представление текста **abaababaabaab**

$X_1 \rightarrow b$
 $X_2 \rightarrow a$
 $X_3 \rightarrow X_2X_1$
 $X_4 \rightarrow X_3X_2$
 $X_5 \rightarrow X_4X_3$
 $X_6 \rightarrow X_5X_4$
 $X_7 \rightarrow X_6X_5$



5 / 11

Результат

Даны две прямолинейные программы, порождающие P и T . Пусть m — количество операций в первой программе, n — во второй. Тогда можно определить, является ли P подстрокой T за время $O(n^2m)$

7 / 11

Поиск сжатой подстроки в сжатом тексте

ВХОД: прямолинейные программы, порождающие P и T

ОТВЕТ: Да/Нет (является ли P подстрокой T ?)

Пример

Text: abaab**aba**abaab

Мы знаем только
сжатое представление
строк P и T

Pattern: **baba**

6 / 11

Разреженная периодичность

Цели:

Найти связь между классической периодичностью и разреженной периодичностью

Построить алгоритм поиска разреженного периода наименьшего размера

8 / 11

Примеры разреженной периодичности

A A B B A A B B C C D D C C D D

Представленная строка неперiodична, но черные клетки

A A B B A A B B C C D D C C D D

образуют структуру, **четырьмя копиями которой** можно замостить исходную строку:

A A B B A A B B C C D D C C D D

Самый простой пример:

A A B B

9 / 11

Основные результаты

- Разработаны полиномиальные алгоритмы поиска сжатых подстрок в сжатых текстах, вычисления минимальных периодов и накрытий сжатого текста, поиска явно заданной подпоследовательности в сжатом тексте
- Доказана $\#P$ -полнота задачи о вычислении расстояния Хэмминга между сжатыми текстами, NP- и coNP-трудность задачи о поиске сжатой подпоследовательности в сжатом тексте
- Показано, что примитивный разреженный период не всегда единственен. Доказано, что каждый примитивный разреженный период “вложен” в любой классический период
- Разработан алгоритм поиска разреженных периодов минимального размера

10 / 11

Спасибо за внимание!

Вопросы?