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.

Тематика работы

1. Вычислительная сложность задач на строках в ситуации, когда на вход алгоритму строки попадают в **заархивированном** виде

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

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 удается, показать NP-трудность рассматриваемых задач

Актуальность

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

Прямолинейные программы

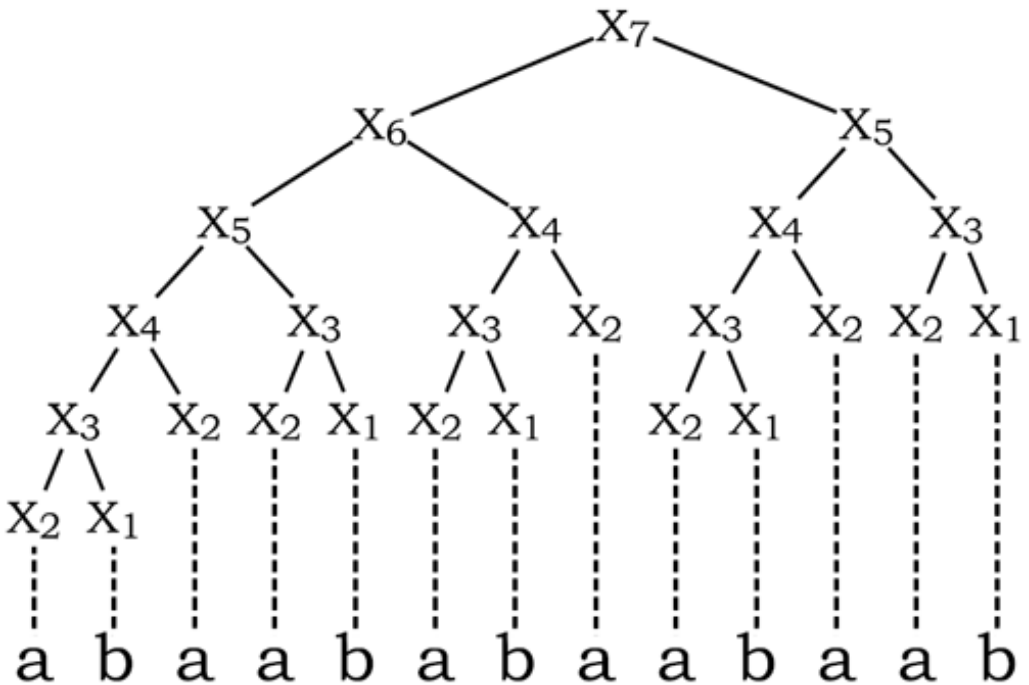
Большинство используемых методов архивирования (семейство Lempel-Ziv, все словарные методы) могут быть эффективно преобразованы в **прямолинейные программы**

Прямолинейные программы

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

Пример: представление текста **abaababaabaab**

$$
\begin{aligned}
X_1 &\to b \\
X_2 &\to a \\
X_3 &\to X_2X_1 \\
X_4 &\to X_3X_2 \\
X_5 &\to X_4X_3 \\
X_6 &\to X_5X_4 \\
X_7 &\to X_6X_5
\end{aligned}
$$



Поиск сжатой подстроки в сжатом тексте

ВХОД: прямолинейные программы, порождающие P и T

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

Пример

Text: abaababaabaab

Pattern: **baba**

Мы знаем только сжатое представление строк P и T

Поиск сжатой подстроки в сжатом тексте

ВХОД: прямолинейные программы, порождающие P и T

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

Пример

Text: abaa**baba**abaab

Pattern: **baba**

Мы знаем только сжатое представление строк P и T

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 $\mathcal{O}(n^2 m)$

Разреженная периодичность

Цели:

Найти связь между классической периодичностью и разреженной периодичностью

Построить алгоритм поиска разреженного периода наименьшего размера

Примеры разреженной периодичности

Представленная строка непериодична, но черные клетки

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

Примеры разреженной периодичности

Представленная строка непериодична, но черные клетки

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

Самый простой пример:

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

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

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

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

Спасибо за внимание!

Вопросы?