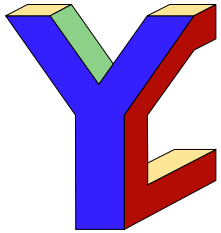


Синтаксический анализ графов для случая стохастических грамматик и вероятностных графов с помощью систем матричных уравнений

Юлия Сусанина

JetBrains Research, Programming Languages and Tools Lab
Санкт-Петербургский Государственный Университет

19.12.2020

Context-Free Path Querying (CFPQ)

- КС-грамматика $G = (N, \Sigma, R)$
  $\mathcal{L}(G_S) = \{\omega \mid S \Rightarrow^*_G \omega\}$, $S \in N$
- Орграф $D = (V, E, \sigma)$, $\sigma \subseteq \Sigma$, $E \subseteq V \times \sigma \times V$
  $m\lambda n$ — путь из m в n в графе D, λ — слово данного пути
- $R_A = \{(m, n) \mid m\lambda n$ — путь в $D, \lambda \in \mathcal{L}(G_A)\}$

CFPQ на языке линейной алгебры

- Через перемножение булевых матриц

---

for all $(i, x, j) \in E$ **do**
$T_{i,j} \leftarrow T_{i,j} \cup \{A \mid (A \rightarrow x) \in P\}$

while матрица T меняется **do**
$T \leftarrow T \cup (T \times T)$

- Как системы матричных уравнений над $\mathbb{R}$

---

for all $N_i \rightarrow \beta_0^0 \dots \beta_k^0 \mid \dots \mid \beta_0^l \dots \beta_m^l$ **do**
решить $\mathcal{T}_{N_i} = \epsilon_{N_i}(T_{\beta_0^0} \cdot \dots \cdot T_{\beta_k^0} + \dots + T_{\beta_0^l} \cdot \dots \cdot T_{\beta_m^l})$,
где ϵ_{N_i}выбрана так, чтобы$\mathcal{T}_{N_i}^k \leq \mathbf{1}$

Мотивация

- Данные из реального мира почти всегда могут быть получены только с некоторой точностью. Детерминированные модели нередко слишком грубо описывают действительность
- Данные могут содержать ошибки и неточности

$$\Downarrow$$

Необходимо находить новые вероятностные модели для более точной обработки существующих массивов данных (например, из биоинформатики)

Stochastic Context-Free Path Querying (SCFPQ)

- Пути с наибольшей вероятностью для всех пар вершин

$$T_A[i,j] == \max_{\substack{A \to BC \\ 0 \le k < n}} \Theta(A \to BC)\, T_B[i,k]\, T_C[k,j]$$

  - Сложность: $(|N^3||V^5|)$

- Вероятности между всеми парами вершин

$$T_A[i,j] = T_A[i,j] + \Theta(A \to BC)\ (T_B T_C)[i,j]$$

or

$$T_{N_i} = \sum_{N_i \to \alpha_j} \Theta(N_i \to \alpha_j)\ T_{\beta_0^j} \cdot \ldots \cdot T_{\beta_k^j}$$

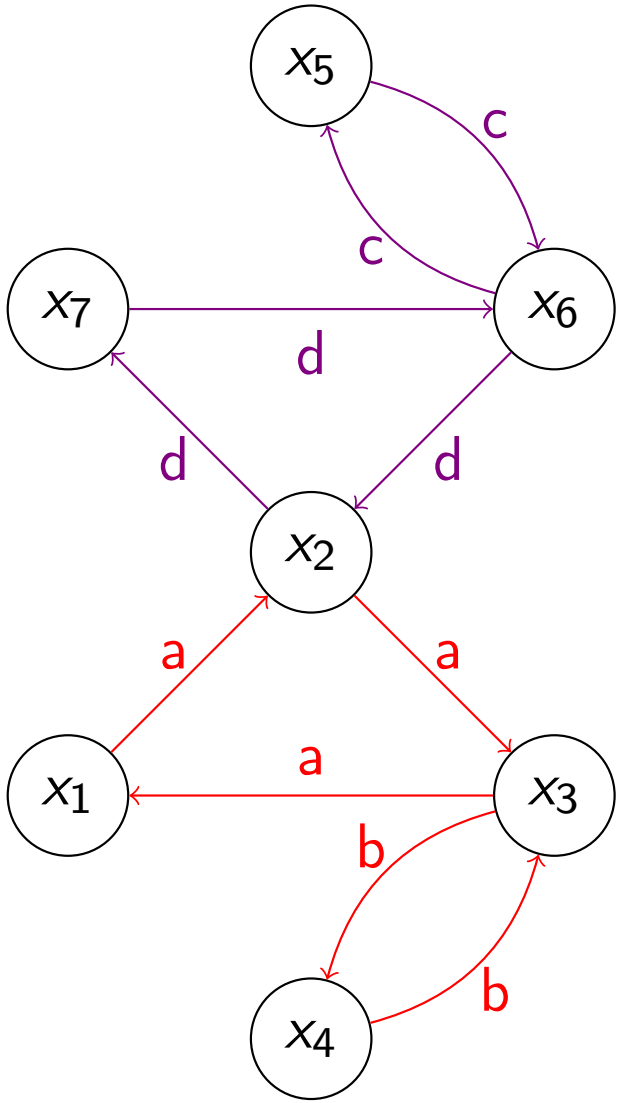
  - В общем случае точное вычисление невозможно

Пример

$$S \xrightarrow{0.3} aSb$$

$$S \xrightarrow{0.6} cSd$$

$$\begin{pmatrix} 0 & 0 & 0.03 & 0.00081 & 0 \\ 0 & 0 & 0.0027 & 0.1 & 0 \\ 0 & 0 & 0.000243 & 0.009 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 & 0 & 0.036 & 0.01296 \\ 0 & 0.0216 & 0 & 0 & 0 & 0.07776 & 0.06 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

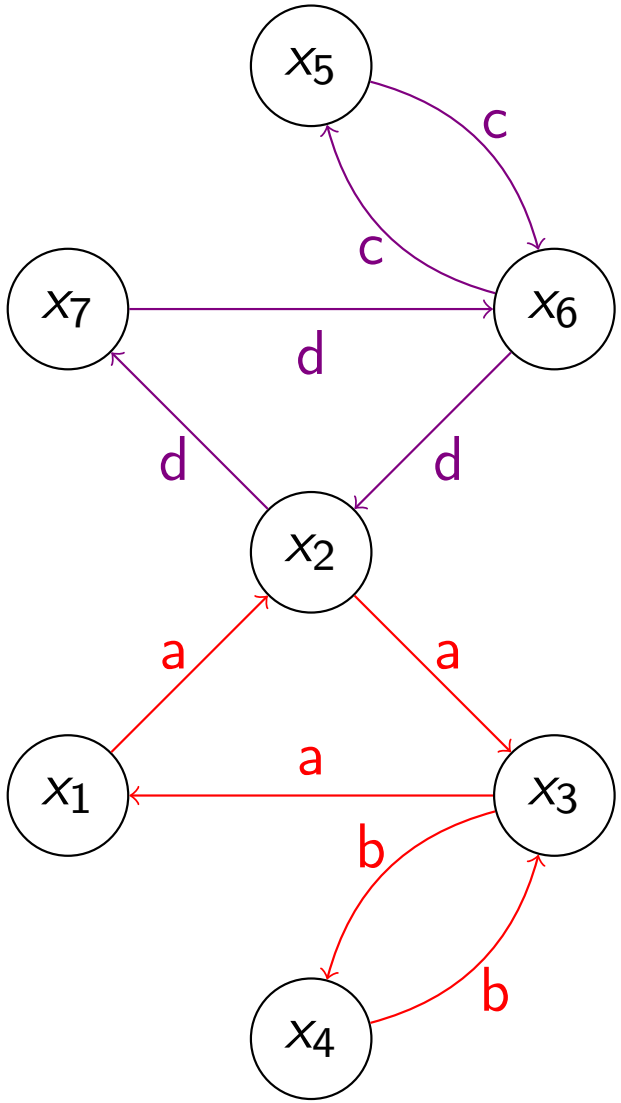


Пример

$$S \xrightarrow{0.3} aSb$$

$$S \xrightarrow{0.6} cSd$$

$$\begin{pmatrix} 0 & 0 & 30000 & 810 & 0 \\ 0 & 0 & 2700 & 100000 & 0 \\ 0 & 0 & 243 & 9000 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 100000 & 0 & 0 & 0 & 36000 & 12960 \\ 0 & 21600 & 0 & 0 & 0 & 7776 & 60000 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$



Взвешенные правила и терминалы

- Веса у правил грамматики — приоритет путям с определенными паттернами
- Веса у вершин графа (или терминалов) — приоритет путям, проходящим через определенные ребра
- Иногда эти два случая совпадают
  - Веса у правил грамматики и у терминалов для однозначных грамматик

Результаты

- Сформулированы две проблемы синтаксического анализа графов для случая стохастических грамматик и вероятностных графов
- Предложены методы решения этих проблем с помощью методов, основанных на алгоритмах линейной алгебры и вычислительной математики
- В процессе
  - Создание эталонного набора данных
  - Эффективная параллельная реализация предложенных алгоритмов с помощью разных итеративных методов и их сравнение
  - Короткая статья на EDBT 2021

Публикации прошлых работ

- **ACM SIGMOD 2020 Student Research Competition:**
  Yuliya Susanina. Context-Free Path Querying via Matrix Equations.
- **LNBI 2020:**
  Yuliya Susanina, Anna Yaveyn and Semyon Grigorev. Modification of Valiant's Parsing Algorithm for String-Searching Problem.
- **Журнал «Труды ИСП РАН»:**
  Сусанина Ю.А., Григорьев С.В. Модификация алгоритма Валианта для задачи поиска подстрок.