Automaty w przetwarzaniu języka naturalnego

Jan Daciuk

Katedra Inteligentnych Systemów Interaktywnych, Wydział ETI, Politechnika Gdańska

HLT Days 2012

Spis treści

1. Dlaczego mnie tu zaproszono?
2. Historia
3. Teraźniejszość
4. Przyszłość

Spis treści

1. Dlaczego mnie tu zaproszono?
2. Historia
3. Teraźniejszość
4. Przyszłość

Dlaczego mnie tu zaproszono?

- konferencje?
- oprogramowanie
  - fsa
  - utr
  - fadd
  - inne
- algorytmy
  - przyrostowego i półprzyrostowego tworzenia minimalnych i pseudominimalnych, deterministycznych automatów skończonych
  - przyrostowego tworzenia automatów drzewiastych
  - tworzenia automatów odgadujących
  - przyrostowej minimalizacji automatów skończonych
  - realizacji funkcji mieszającej z użyciem automatów drzewiastych

Oprogramowanie – pakiet fsa

- zestaw programów:
  - fsa_build
  - fsa_ubuild
  - fsa_accent
  - fsa_guess
  - fsa_hash
  - fsa_morph
  - fsa_prefix
  - fsa_spell
  - fsa_synth
  - fsa_visual
- minimalne automaty skończone poddane znacznej kompresji
- programy w znacznym stopniu niezależne od języka

Oprogramowanie – pakiet utr

- zestaw programów:
  - tr_ubuild
  - tr_accent
  - tr_morph
  - tr_prefix
  - tr_spell
- minimalne automaty Mealy'ego ze znaczną kompresją
- programy w znacznym stopniu niezależne od języka

Oprogramowanie – inne

- **fadd** — wykorzystywanie automatów skończonych do poprawiania pisowni, dopisywania znaków diakrytycznych, analizy i syntezy morfologicznej, indeksowania, modeli języka (biblioteka w C++ z interfejsem C kompatybilna z pakietem fsa)
- zestawy programów realizujących różne algorytmy:
  - przyrostowego i półprzyrostowego tworzenia automatów minimalnych i pseudominimalnych
  - przyrostowego tworzenia automatów drzewiastych
  - minimalizacji automatów skończonych
  - funkcji mieszającej z użyciem automatów drzewiastych (minimalnych i pseudominimalnych)
  - innych...

Spis treści

1. Dlaczego mnie tu zaproszono?
2. Historia
3. Teraźniejszość
4. Przyszłość

ISSCO

54
BUTLER MARINE S.A.
UNION UAP Associates
GLOBAL DATA S.A.
PROSTAR LTD
AUTOLAND S.A.

Archamps

ISSCO

FSMNLP

FSMNLP 2001

Alicante

Spis treści

1. Dlaczego mnie tu zaproszono?
2. Historia
3. Teraźniejszość
4. Przyszłość

Teraźniejszość: kompresja automatów?

- artykuł na konferencji CIAA 2011 w Blois
  - nagrodzony jako najlepszy (jeden z dwóch)
- artykuł w Theoretical Computer Science (wrzesień 2012, rozszerzona wersja artykułu z konferencji)
- pomysł, programowanie i doświadczenia — Dawid Weiss

Teraźniejszość: nic nie robię (naukowo)...

...a dokładniej: piszę rozprawę habilitacyjną

Incrementality and Hashing in Automata

Jan Daciuk
Department of Intelligent Information Systems
Faculty of Electronics, Telecommunications and Informatics
Gdańsk University of Technology

September 18, 2012

Spis treści

1. Dlaczego mnie tu zaproszono?
2. Historia
3. Teraźniejszość
4. Przyszłość

Poprawki i rozszerzenia istniejących pakietów

(Głównie fsa i fadd):

- poprawki zgłoszone przez Marcina Miłkowskiego
- UTF8 (poprawianie pisowni i dopisywanie znaków diakrytycznych)
- nowa kompresja (Dawid Weiss)
- morfologia derywacyjna

Poprawki oprogramowania i algorytmu

Algorytm przyrostowej minimalizacji opublikowany z Brucem Watsonem zawiera błąd:

- błąd został odkryty 2 lata temu przez Marco Almeidę
- poprawiłem błąd, ale Bruce nie współpracuje przy artykule
- Marco opublikował własną wersję algorytmu, według niego szybszą
- sądzę, że moja wersja może jednak być szybsza (po poprawkach)

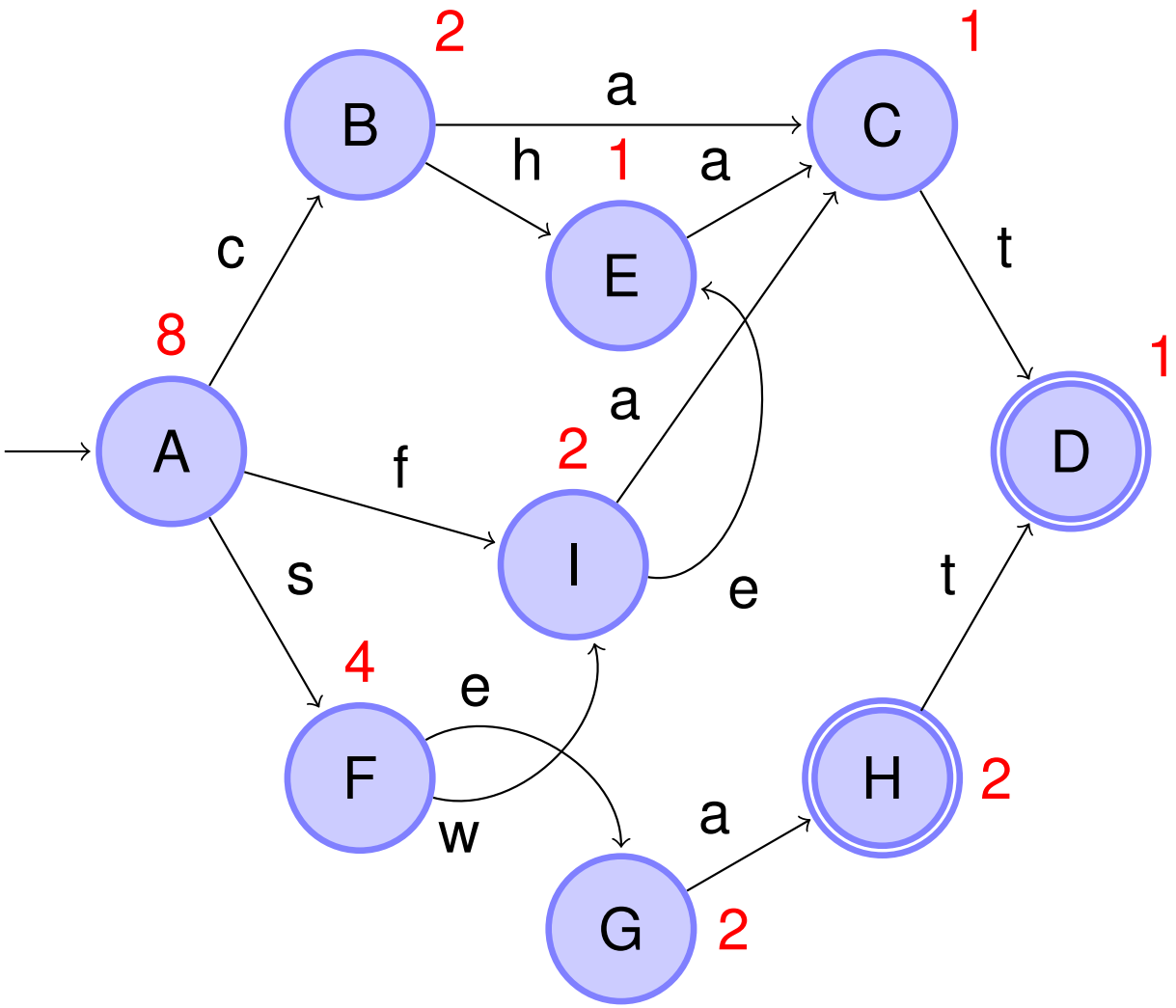
Kompresja oznaczonych zbiorów tekstów

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<alpino_ds version="1.0">
 <node id="0" rel="top" cat="top" begin="0" end="26">
  <node id="1" rel="–" cat="smain" begin="0" end="25">
   <node id="2" rel="vc" cat="inf" begin="0" end="25">
    <node id="3" rel="vc" cat="inf" begin="0" end="25">
     <node id="4" rel="mod" cat="pp" begin="0" end="3">
      <node id="5" rel="hd" pos="prep" begin="0" end="1" root="in" word="In"/>
      <node id="6" rel="obj1" cat="np" begin="1" end="3">
       <node id="7" rel="mod" pos="adj" begin="1" end="2" root="uitzonderlijk" word="uitzonderlijke"/>
       <node id="8" rel="hd" pos="noun" begin="2" end="3" root="geval" word="gevallen"/>
      </node>
     </node>
     <node id="9" rel="su" index="1" cat="np" begin="4" end="6">
      <node id="10" rel="det" pos="det" begin="4" end="5" root="de" word="de"/>
      <node id="11" rel="hd" pos="noun" begin="5" end="6" root="regering" word="regering"/>
     </node>
```

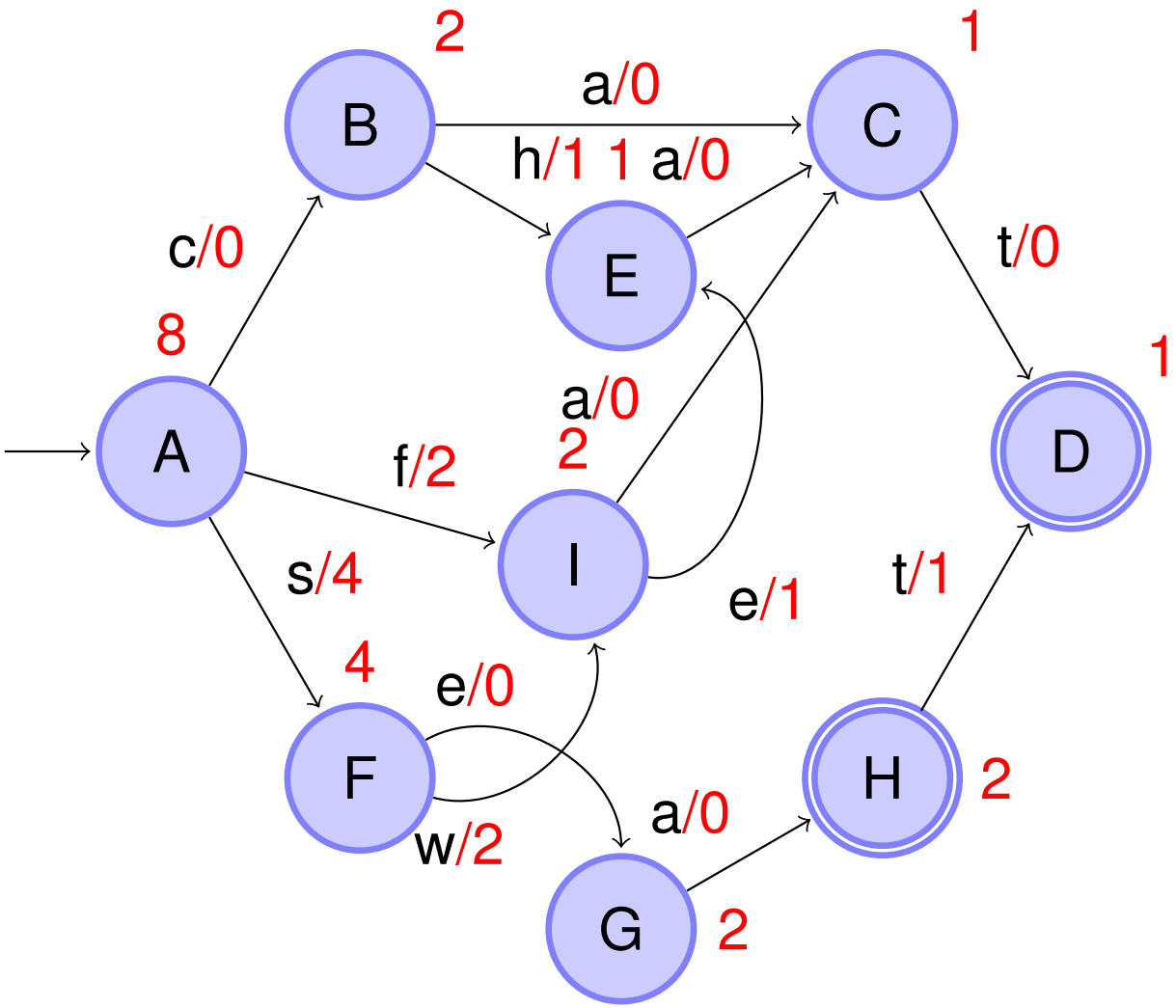
Złożenie kilku elementów:

- automatów skończonych z funkcją mieszającą
- automatów drzewiastych z funkcją mieszającą
- wektorów indeksów
- wektorów dodatkowych informacji

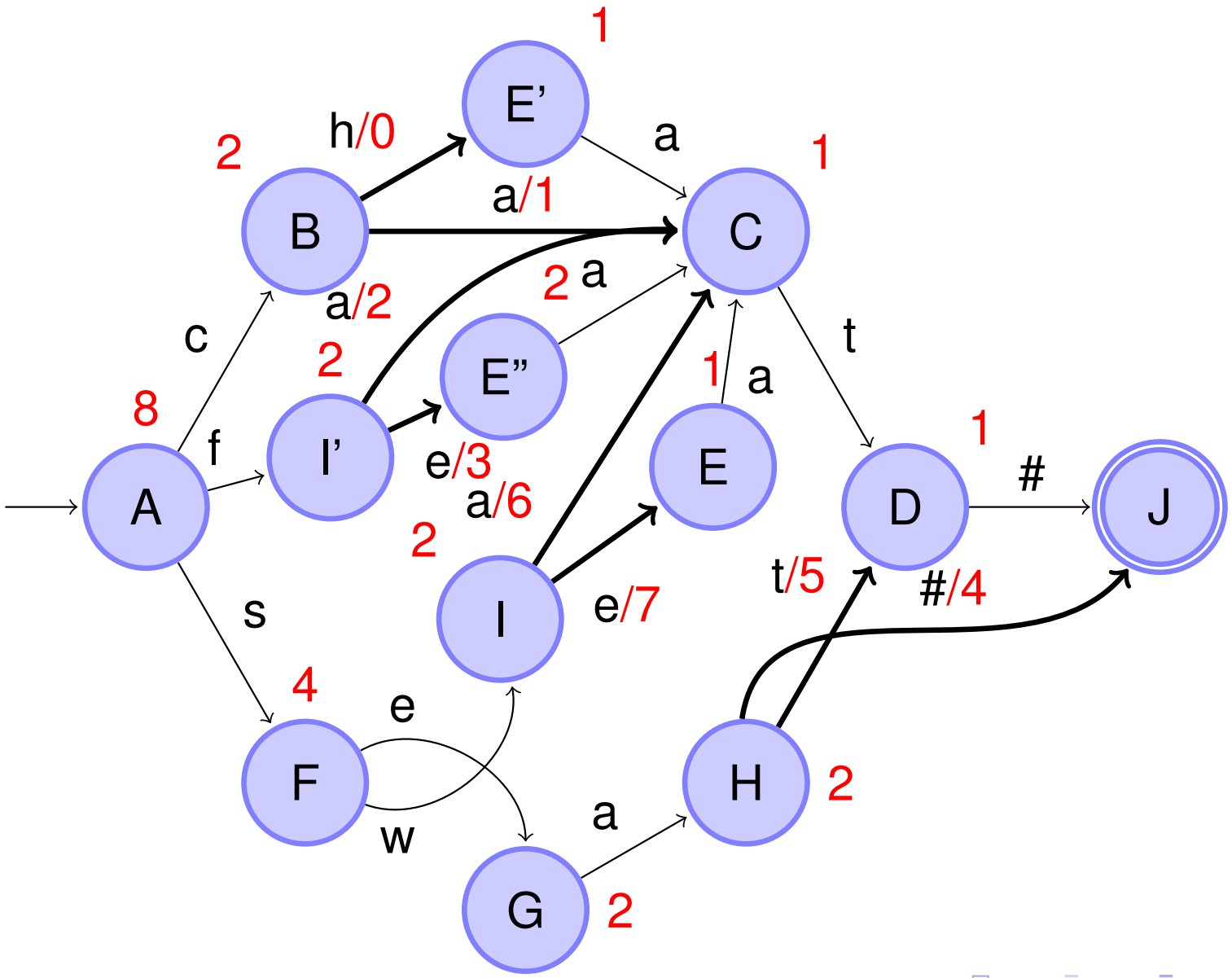
Automaty skończone z funkcją mieszającą



Automaty skończone z funkcją mieszającą



Automaty pseudominimalne z funkcją mieszającą

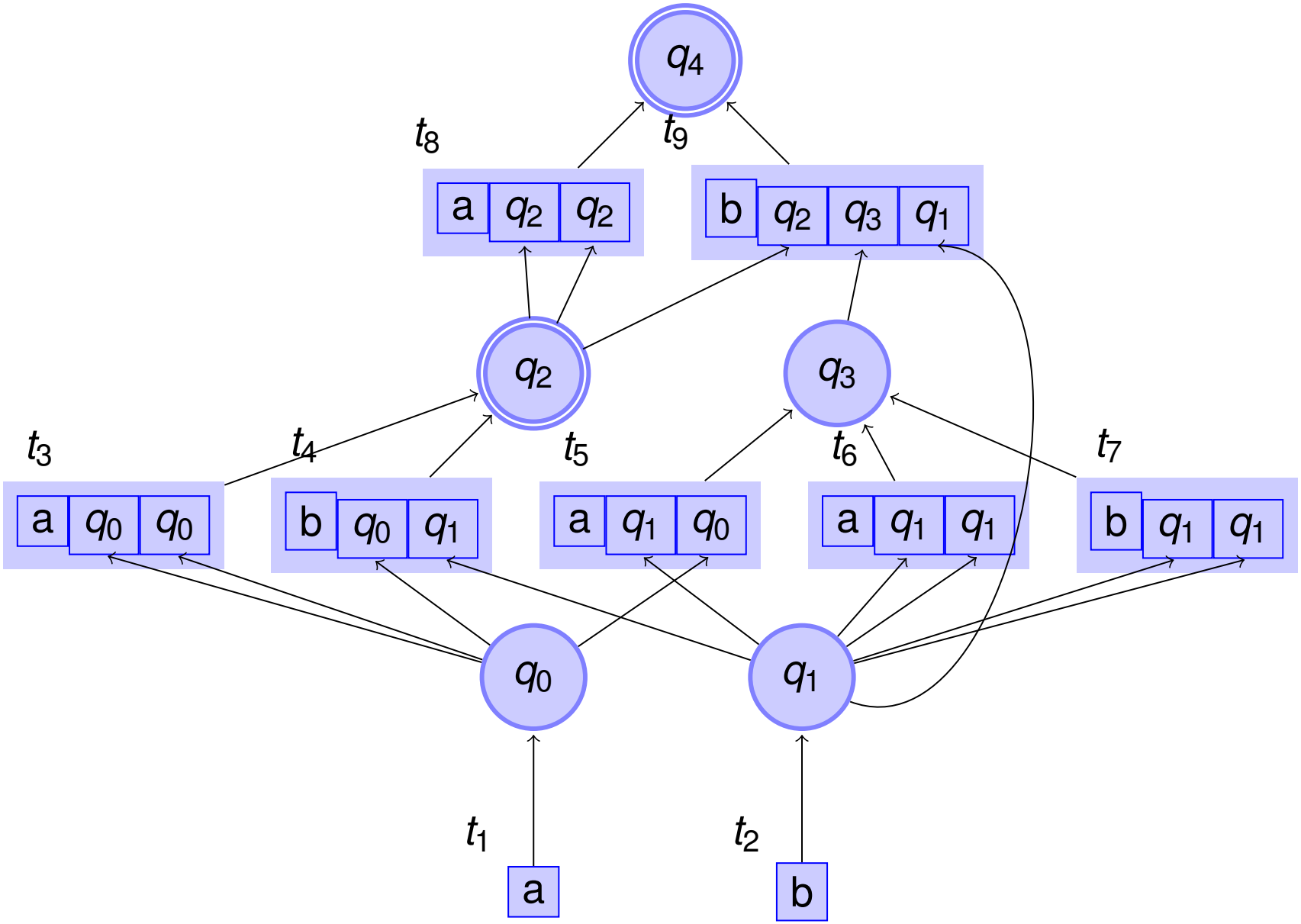


Automaty drzewiaste z funkcją mieszającą

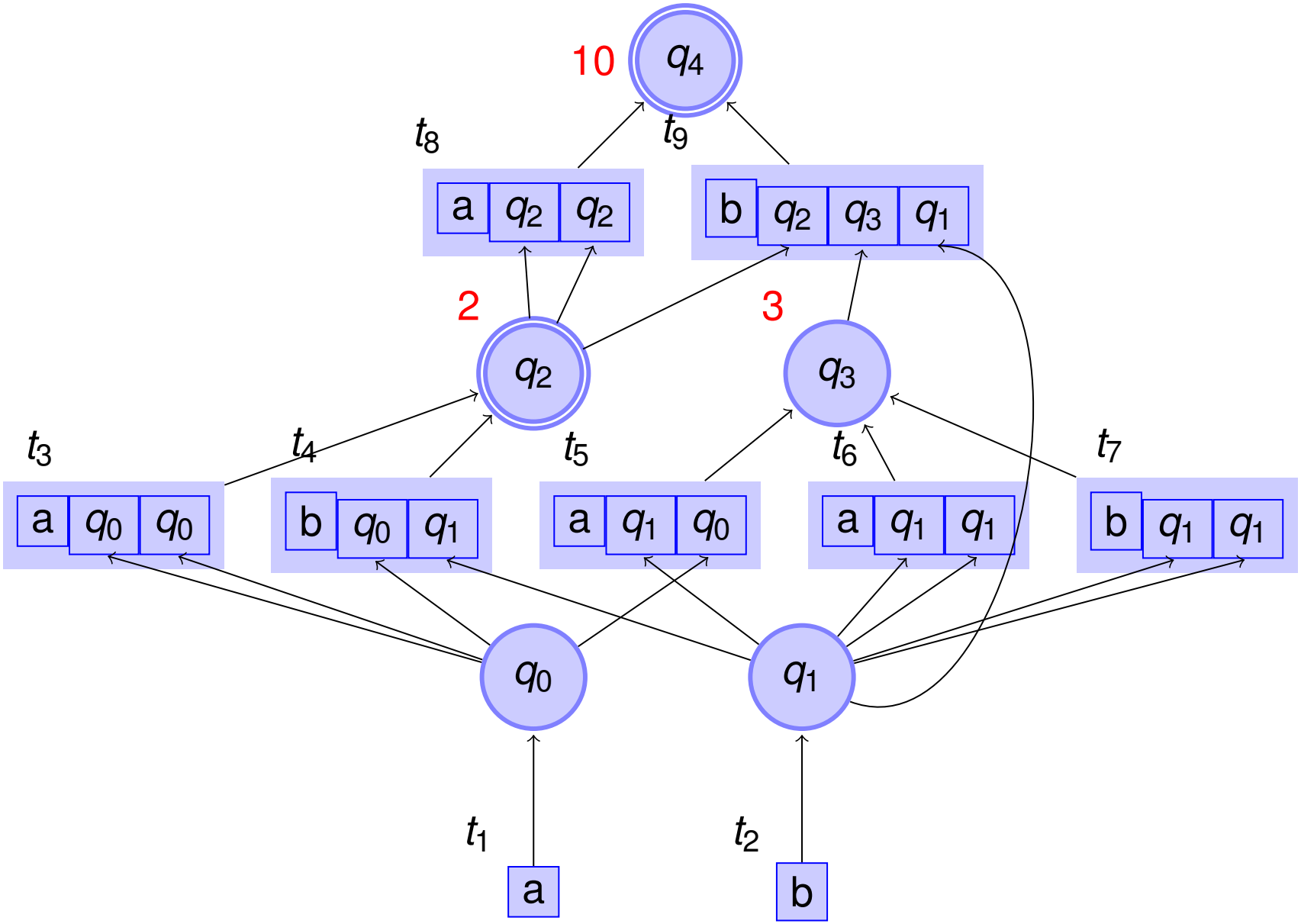
Automaty drzewiaste:

- rozpoznają drzewa
- nadają się do rozpoznawania struktur XML, rozbiorów gramatycznych itp.
- występują w dwóch wersjach:
  - wstępującej: rozpoznanie od liści do korzenia
  - zstępującej: od korzenia do liści
- w wersji wstępującej: $\delta_n : \Sigma \times Q^n \rightarrow Q, n \geq 0$, brak stanu początkowego

Automaty drzewiaste z funkcją mieszającą



Automaty drzewiaste z funkcją mieszającą



Kompresja oznaczonych zbiorów tekstów

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<alpino_ds version="1.0">
 <node id="0" rel="top" cat="top" begin="0" end="26">
  <node id="1" rel="–" cat="smain" begin="0" end="25">
   <node id="2" rel="vc" cat="inf" begin="0" end="25">
    <node id="3" rel="vc" cat="inf" begin="0" end="25">
     <node id="4" rel="mod" cat="pp" begin="0" end="3">
      <node id="5" rel="hd" pos="prep" begin="0" end="1" root="in" word="In"/>
      <node id="6" rel="obj1" cat="np" begin="1" end="3">
       <node id="7" rel="mod" pos="adj" begin="1" end="2" root="uitzonderlijk"
         word="uitzonderlijke"/>
       <node id="8" rel="hd" pos="noun" begin="2" end="3" root="geval" word="gevallen"/>
      </node>
     </node>
     <node id="9" rel="su" index="1" cat="np" begin="4" end="6">
      <node id="10" rel="det" pos="det" begin="4" end="5" root="de" word="de"/>
      <node id="11" rel="hd" pos="noun" begin="5" end="6" root="regering" word="regering"/>
     </node>
```

Co powinno być etykietą przejścia/wierzchołka drzewa?

Kompresja oznaczonych zbiorów tekstów

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<alpino_ds version="1.0">
 <node id="0" rel="top" cat="top" begin="0" end="26">
  <node id="1" rel="–" cat="smain" begin="0" end="25">
   <node id="2" rel="vc" cat="inf" begin="0" end="25">
    <node id="3" rel="vc" cat="inf" begin="0" end="25">
     <node id="4" rel="mod" cat="pp" begin="0" end="3">
      <node id="5" rel="hd" pos="prep" begin="0" end="1" root="in" word="In"/>
      <node id="6" rel="obj1" cat="np" begin="1" end="3">
       <node id="7" rel="mod" pos="adj" begin="1" end="2" root="uitzonderlijk"
          word="uitzonderlijke"/>
       <node id="8" rel="hd" pos="noun" begin="2" end="3" root="geval" word="gevallen"/>
      </node>
     </node>
     <node id="9" rel="su" index="1" cat="np" begin="4" end="6">
      <node id="10" rel="det" pos="det" begin="4" end="5" root="de" word="de"/>
      <node id="11" rel="hd" pos="noun" begin="5" end="6" root="regering" word="regering"/>
     </node>
```

Co powinno być przechowywane?

Kompresja oznaczonych zbiorów tekstów

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<alpino_ds version="1.0">
 <node id="0" rel="top" cat="top" begin="0" end="26">
  <node id="1" rel="–" cat="smain" begin="0" end="25">
   <node id="2" rel="vc" cat="inf" begin="0" end="25">
    <node id="3" rel="vc" cat="inf" begin="0" end="25">
     <node id="4" rel="mod" cat="pp" begin="0" end="3">
      <node id="5" rel="hd" pos="prep" begin="0" end="1" root="in" word="In"/>
      <node id="6" rel="obj1" cat="np" begin="1" end="3">
       <node id="7" rel="mod" pos="adj" begin="1" end="2" root="uitzonderlijk"
          word="uitzonderlijke"/>
       <node id="8" rel="hd" pos="noun" begin="2" end="3" root="geval" word="gevallen"/>
      </node>
     </node>
     <node id="9" rel="su" index="1" cat="np" begin="4" end="6">
      <node id="10" rel="det" pos="det" begin="4" end="5" root="de" word="de"/>
      <node id="11" rel="hd" pos="noun" begin="5" end="6" root="regering" word="regering"/>
     </node>
```

Co powinno być przechowywane?

Kompresja oznaczonych zbiorów tekstów

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<alpino_ds version="1.0">
 <node id="0" rel="top" cat="top" begin="0" end="26">
  <node id="1" rel="–" cat="smain" begin="0" end="25">
   <node id="2" rel="vc" cat="inf" begin="0" end="25">
    <node id="3" rel="vc" cat="inf" begin="0" end="25">
     <node id="4" rel="mod" cat="pp" begin="0" end="3">
      <node id="5" rel="hd" pos="prep" begin="0" end="1" root="in" word="In"/>
      <node id="6" rel="obj1" cat="np" begin="1" end="3">
       <node id="7" rel="mod" pos="adj" begin="1" end="2" root="uitzonderlijk"
         word="uitzonderlijke"/>
       <node id="8" rel="hd" pos="noun" begin="2" end="3" root="geval" word="gevallen"/>
      </node>
     </node>
     <node id="9" rel="su" index="1" cat="np" begin="4" end="6">
      <node id="10" rel="det" pos="det" begin="4" end="5" root="de" word="de"/>
      <node id="11" rel="hd" pos="noun" begin="5" end="6" root="regering" word="regering"/>
     </node>
```

Co powinno być przechowywane?