Rámec na učenie vzorov z dokumentov

Bc. Miroslav Legéň

Vedúci diplomového projektu: Mgr. György Frivolt

Slovenská technická univerzita v Bratislave

Fakulta informatiky a informačných technológií

Motivácia projektu

- Rôzne reprezentácie dokumentov – html,txt,doc
- Neznalosť jazykov XPath, HTML
- Ručné vytváranie obaľovačov
- Využitie strojového učenia pri tvorbe a získavaní vzorov z dokumentov

Obaľovače webových stránok

- Dokument reprezentovaný DOM stromom vs plain text
- DEByE, XWRap, Lixto, Wrapper Suite
- Dokument – množina symbolov
  - Regulárne výrazy
  - HMM
  - MEM
- Bibliografické odkazy

Skrytý markov model - HMM

- Systémy, ktorých vnútorný stav nie je známy
- Množina viditeľných symbolov previazaná s vnútornými stavmi modelu
- Prechody medzi stavmi
- Pravdepodobnostné rozdelenie emisie symbolu
- Vytvorenie modelu - učenie
  - sekvencia pozorovaní
  - učíme model pomocou Baum-welch algoritmu

Skrytý markov model - HMM

- Dekódovanie
  - Parametre modelu
  - Sekvencie pozorovaní
  - Hľadáme cestu – prechod stavmi – Viterbiho algoritmus
- Vyhodnotenie
  - Parametre modelu
  - Sekvencie pozorovaní
  - Výber najvhodnejšieho modelu pre sekvenciu symbolov Forward-Backward algoritmus

Skrytý markov model

- Výhody
  - Transparentnosť, výkonnosť
  - Učenie a spájanie modelov
  - Informácie z kontextu
- Nevýhody
  - Závislosť modelu od trénovacích vzoriek

Použitie:

- Rozpoznávanie reči, OCR, prekladanie, POS, bioinformatika

Vzorom je model

- Informácie môžu byť v pološtruktúrovanej forme

Bellman, R. 1957. Dynamic Programming. Princeton University Press.

- Dokument – množina symbolov so skrytou štruktúrou
- Objavenie skrytej štruktúry pomocou HMM
- Pozitívne,negatívne príklady - flexibilita
- HMM predstavuje vzor pre extrakciu informácii z dokumentu

Metóda pre tvorbu HMM

- Vyznačenie relevantných štítkov
  - časti textu pre extrakciu
  - stavy HMM
  - stavy modelu tvoria kliku
- Prechody -> následnosť štítkov
- Slová rozdelené do skupín
- Učenie: Baum–Welch, K-Means algoritmy
- Aplikovanie modelu pomocou Viterbiho algoritmu

Plán na ďalší semester

- Doplňujúce sémantické informácie zo slovníka – WORDNET
- Optimalizácia BW algoritmu pre PN príklady
- GUI pre umožnenie interakcie
- Integrácia do Wrapper Suite
- Použitie aj iných typov HMM
  - hierarchické
  - vrstvové

Ďakujem za pozornosť.