



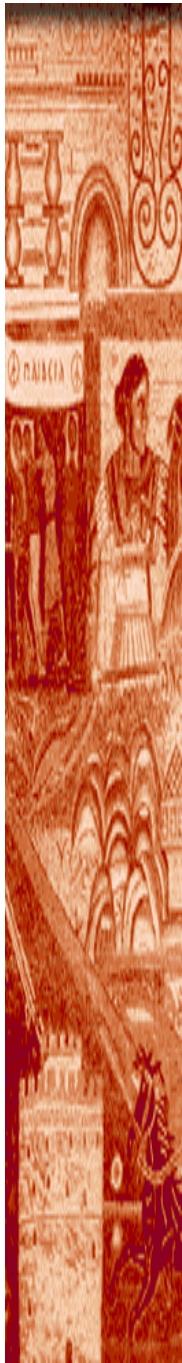
Automatické sémantické anotování a učení ontologií

Vojtěch Svátek

Vysoká škola ekonomická v Praze
katedra informačního a znalostního inženýrství
svatek@vse.cz

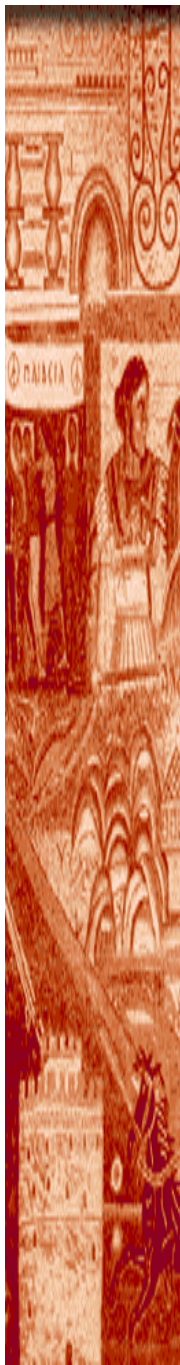
Sémantický web a textové zdroje

- Sémantický web je (primárně) určen pro softwarové aplikace – hlavní je pro něj *formálně strukturovaná* reprezentace
- Podstatou současného webu jsou převážně *texty* (v menší míře obrázky) v *prezentační* struktuře (HTML)
- Pro vznik „nadkritického“ množství formálně strukturovaných („sémantických“) dat je nezbytné využít existující texty a prezentační strukturu



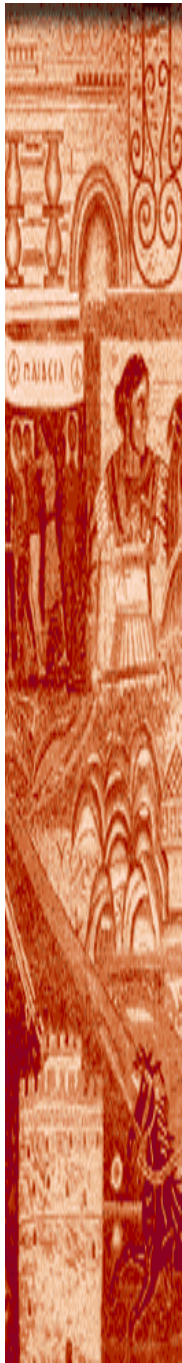
Sémantický web a textové zdroje (2)

- Transformace textu na sémantické struktury (např. RDF) pomocí vyznačování jeho částí se označuje jako *sémantické anotování*
 - ruční
 - poloautomatické
 - automatické
- *Automatické* anotování je založené na metodách označovaných jako *extrakce informací* (information extraction – IE)



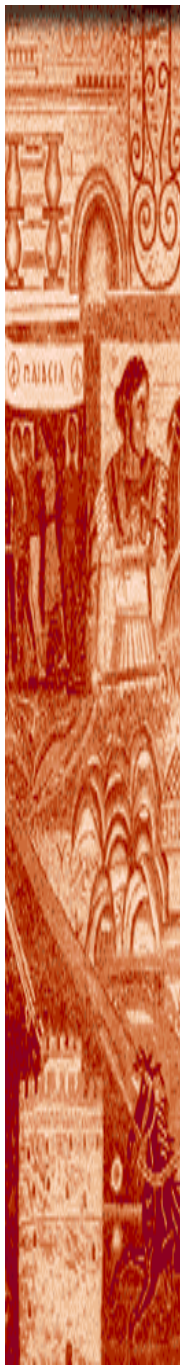
Sémantický web a textové zdroje (3)

- Ontologie jsou obvykle méně rozsáhlé a stabilnější než báze RDF faktů
- I tak je ale jejich tvorba náročná a je obtížné dosáhnout reprezentativního pokrytí problémové oblasti
- Automatickou analýzou (dolováním z) textů lze nalézt
 - termíny – kandidáty na třídy, relace a instance
 - taxonomické a netaxonomické vztahy
 - někdy i další logické axiomy
- Tento proces se často označuje jako *učení ontologií*



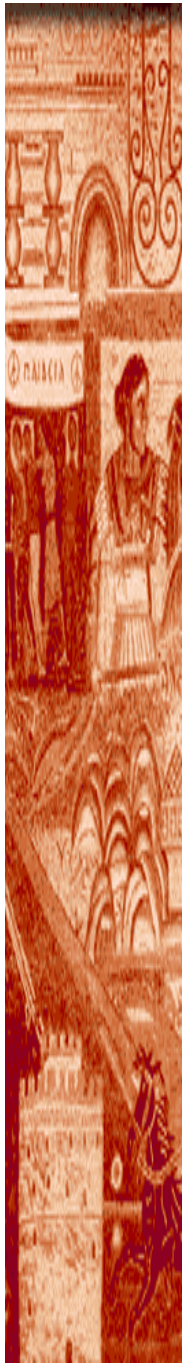
Extrakce informací

- Prehistorie již několik desítek let v rámci strojové lingvistiky – sémantická analýza struktury vět
 - nadstavba plné syntaktické větné analýzy
 - snaha o preciznost a obecnost (nezávislost na doméně)
 - náročné ruční anotování dat, nízká adaptovatelnost pro specifickou doménu
 - dnes např. tzv. tektogramatická vrstva pražského závislostního korpusu



Extrakce informací (2)

- „Pragmatická“ větev IE vznikla koncem 80. let jako prostředek pro rychlé vyhledávání klíčových informací v krátkých textových zprávách, např.
 - nehody, teroristické/kriminální činy...
 - obchodní svět (akvizice, personální změny)
- Brzy rozšíření do dalších oblastí, např.
 - předpovědi počasí
 - lékařské zprávy
- ... a obecně pro webová data: *web IE*

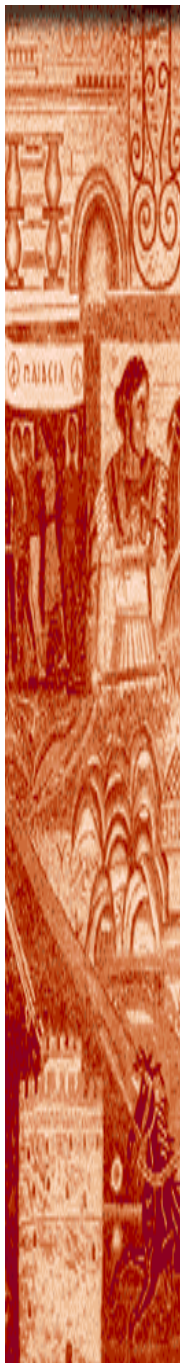


Extrakce informací (3)

- Zpočátku většinou založené na jednoduchých ručně formulovaných vzorech (vzorcích?) – *regulární výrazy*
- Příklad z oblasti medicíny – extrakce hodnot krevního tlaku
$$TK ([0-9]+) / ([0-9]+)$$
- Na rozdíl od „čistého“ lingvistického přístupu funguje i pro „útržkovitý“ text

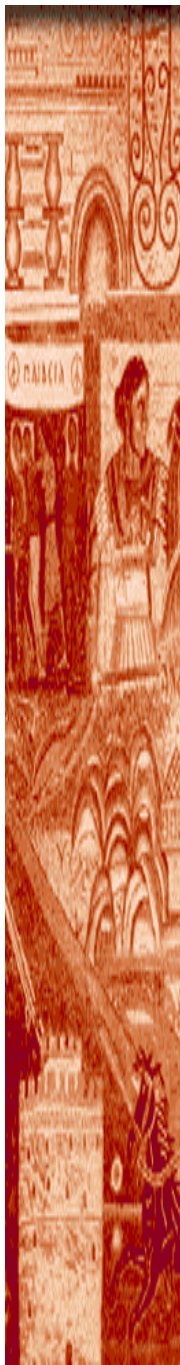
Extrakce informací (4)

- Ruční tvorba vzorů je často subjektivně ovlivněná a při nárůstu jejich počtu je obtížně je udržovat
- Hlavním přístupem se později stalo *učení* vzorů, ať už v rámci
 - symbolických pravidel (explicitní vzory)
 - statistických modelů (implicitní vzory skryté v pravděpodobnostních distribucích)
 - wrapperů (explicitní vzory nad elementy HTML)
- Podrobněji viz prezentace M. Labského



Extrakce informací (5)

- Učení vzorů ovšem vyžaduje ručně anotovaná trénovací data/příklady
- Wrappery
 - stačí několik málo příkladů, ale omezené využití (závislost na strukturovanosti stránky)
- Pravidla
 - větší množství trénovacích dat
- Statistické modely
 - velké množství trénovacích dat

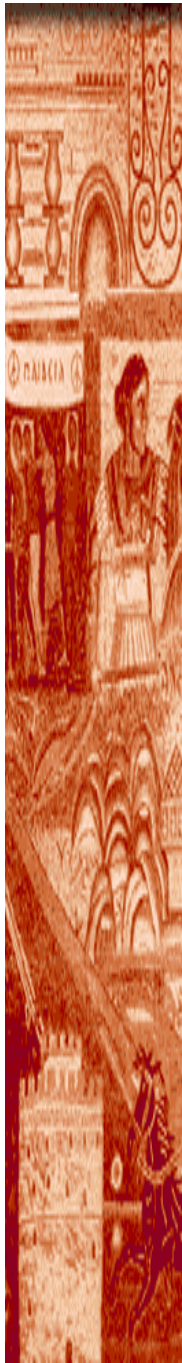


Extrakce informací (6)

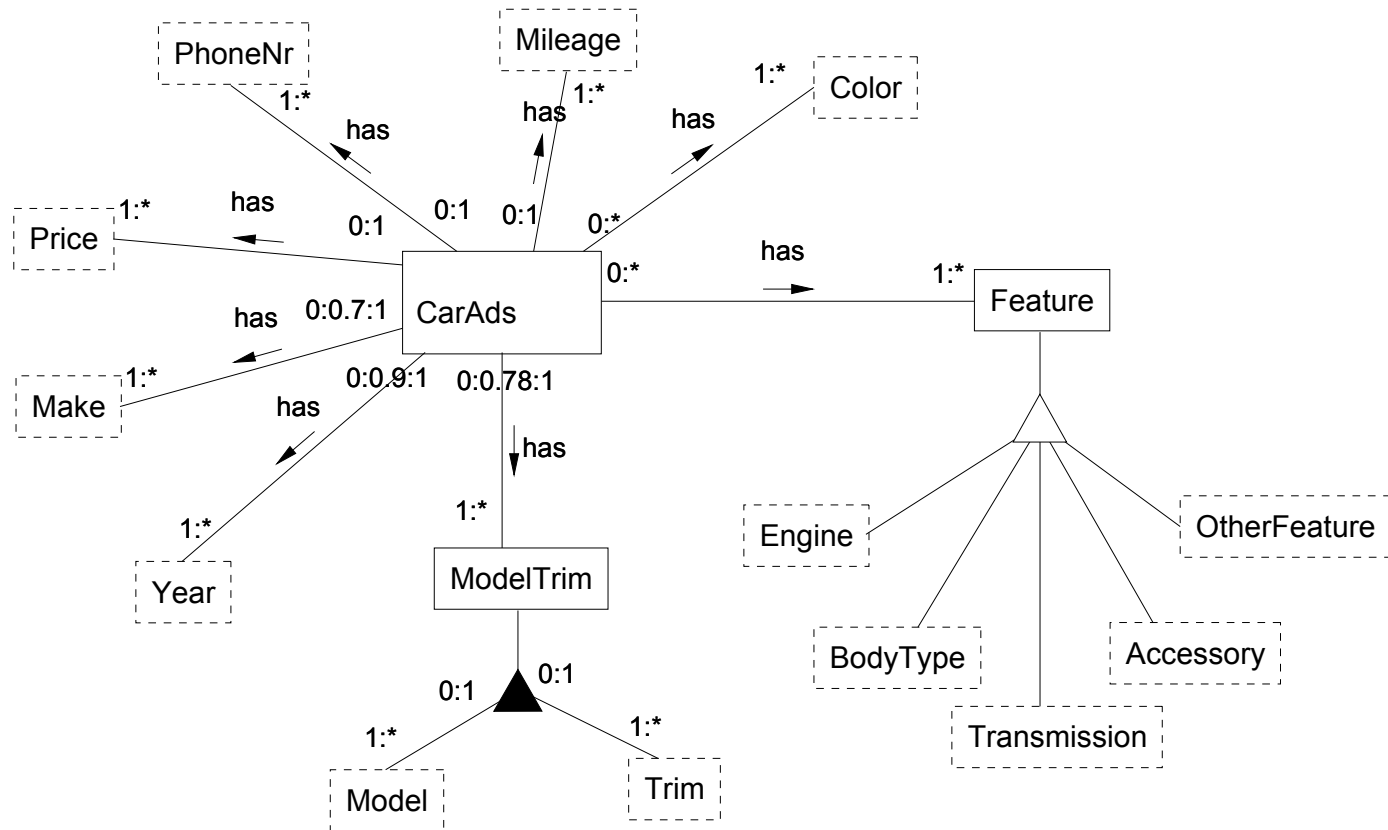
- Čistě ruční tvorba trénovacích dat je velmi nákladná, proto se používají iterativní procesy
 - *Statistický bootstrapping*: vzory, které jsou velmi úspěšné na malém vzorku ručně anotovaných trénovacích dat jsou následně použity pro anotování dalších dat (nese s sebou riziko propagace chyb)
 - Bootstrapping založený na *redundanci* informací (zejména pro WWW): z informace, kterou systém najde na různých zdrojích v různé struktuře, odvodí formální tvar informací v těchto zdrojích (např. biblio, inzeráty – systém Armadillo) a podle toho z nich extrahuje informace o dosud neznámých objektech

Extrakce informací (7)

- Vedle toho se stále uplatňují přístupy založené na ruční tvorbě vzorů (zpravidla v kombinaci s učením ev. wrappery)
- Perspektivní jsou zejména přístupy založené na *extrakčních ontologiích* (Embley, Labský)
- Výhoda rychlého startu – vytvoří se zárodek modelu, který je iterativně vylepšován
- Souvislost mezi extrakčními a „normálními“ doménovými ontologiemi – možnost částečné transformace jedněch na druhé



Extrakční ontologie pro nabídky aut (Embley 2006)



Extrakční ontologie pro nabídky aut (Embley 2006)

```

<ObjectSet x="329" y="51" lexical="true" name="Mileage" id="osmx50">
  <DataFrame>
    <InternalRepresentation>
      <DataType typeName="String"/>
    </InternalRepresentation>
    <ValuePhraseList>
      <ValuePhrase hint="Mileage Pattern 1">
        <ValueExpression color="ffffff">
          <ExpressionText>[1-9]\d{0,2}[kK]</ExpressionText>
        </ValueExpression>
        <LeftContextExpression color="ffffff">
          ...
        </LeftContextExpression>
      </ValuePhrase>
      <KeywordPhraseList>
        <KeywordPhrase hint="New phrase 1">
          <KeywordExpression color="ffffff">
            <ExpressionText>\bmiles\b</ExpressionText>
          </KeywordExpression>
          ...
        </KeywordPhrase>
      </KeywordPhraseList>
    </ValuePhraseList>
  </DataFrame>
</ObjectSet>

```

Část extrakční ontologie pro kontaktní informace (Labský 2007)

```
<class id="Contact">
```

```
...
```

```
<attribute id="street" type="name" card="0-1" eng="0.50">
```

```
<pattern id="street_suffix_cap" case="CA|UC"> Way | Lake | Square | Place  
Ridge | Beach | Valley | Ridge </pattern>
```

```
<pattern id="street_suffix" ignore="case">
```

```
street | st .? | road | rd .? | parkway | pkwy .? | drive | drv .? | sq .?
```

```
boulevard | boul .? | blv .? | blvd .? | pl .? | pike | turnpike | turn pike
```

```
avenue | ave .? | av .? | lane | ln . | lk .
```

```
<pattern ref="street_suffix_cap"/> </pattern>
```

```
<pattern id="geo_dirs">
```

```
Northern | Southern | Western | Eastern | North | South | West | East
```

```
( (NW | NE | SW | SE | E | W | N | S) .? ) </pattern>
```

```
<pattern id="full_street_pattern">
```

```
<tok type="INT"/>? ( ( <tok type="INT"/> ( (-|.)? (th|rd|st|nd) )? ) |
```

```
<tok type="NUMALPHA"/>)
```

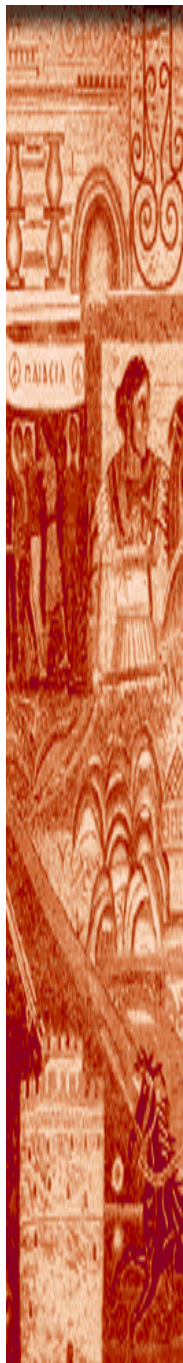
```
<tok type="ALPHA|NUMALPHA" case="CA|UC"/>{0,3}
```

```
<pattern ref="street_suffix"/> ( ,? <pattern ref="geo_dirs"/> )? </pattern>
```

```
...
```

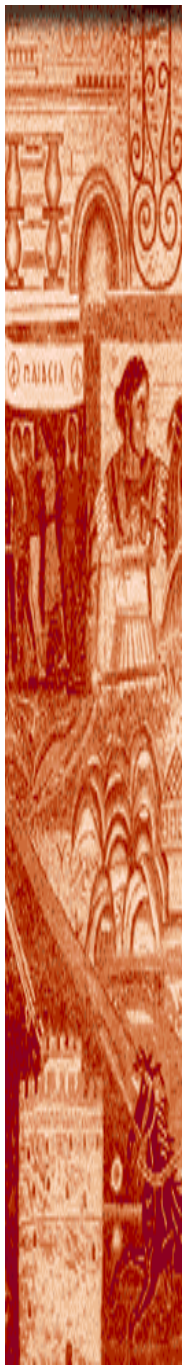
```
</attribute>
```

```
...
```



Učení ontologií

- Víceméně kopíruje proces ruční tvorby ontologií, ale snaží se využít automatické techniky
- Hlavní fáze
 - extrakce klíčových termínů
 - identifikace tříd a instancí pojmů
 - tvorba taxonomie
 - tvorba a pojmenování netaxonomických relací
 - tvorba složitějších axiomů, a charakterizace ve smyslu „upper-level“ (např. „látkové“ pojmy...)



Učení ontologií (2)

- Dva hlavní směry (často se prolínají)
 - směr založený na četnostech termínů v dokumentech, např.
 - pokud ve většině dokumentů, kde se vyskytuje t2, se také (lépe: v jeho blízkosti) vyskytuje t1, pak by t2 mohl označovat podtřídu vzhledem k t1
 - pokud se t1 a t2 vyskytují ve většině dokumentů v blízkosti jeden druhého, mohlo by jít o netaxonomickou relaci
 - směr založený na strukturních vzorech (Hearst patterns) – souvislost s IE
 - např.: „X a jiné Y“, „X je Y, který...“, „...tyto Y: X, ...“