EXTRAKCE STRUKTUROVANÝCH DAT O PRODUKTOVÝCH A PRACOVNÍCH NABÍDKÁCH POMOCÍ EXTRAKČNÍCH ONTOLOGIÍ

ALEŠ POUZAR

PŘEDMĚT PRÁCE

Popis

- extrakce strukturovaných dat ve vybraných doménách ze semistrukturovaných webových stránek nezávisle na jejich struktuře
- předmětem extrakce jsou
  - katalogové a produktové stránky eshopů (český web)
  - firemní stránky (český web)
- experimenty byly provedeny s extrakčním systémem Ex

Motivace

- získat data o vysoké granularitě, která by mohla být využitelná v praktických aplikacích (srovnávač produktových nabídek, vyhledávač pracovních pozic)

VYMEZENÍ V RÁMCI IE

Oblast

- web content mining
- web structure mining
- web usage mining

Typy úloh

- extrakce pojmenovaných entit (named entity extraction)
- extrakce relací (relation extraction)

Kombinace přístupů

- ručně zadaná pravidla
- indukce wrapperů
- učící algoritmy
- NLP techniky

Obecnost metod

- homogenní skupina dokumentů
- omezení na doménu
- bez omezení

KLASIFIKACE WIE SYSTÉMŮ

Vymezení systému Ex v rámci WIE

- úroveň strukturovanosti dokumentu:
  structured, semi-structured, free text
- úroveň extrakce:
  field-level, record-level, page-level, site-level
- tokenizace: word-level vs. tag-level
- seskupování atributů do instancí: top-down vs. bottom-up
- variace extrakčního cíle
  - chybějící atributy
  - mnohonásobný výskyt atributu
  - permutace atributů
  - vnořené objekty (podporováno nepřímo)
- typy extrakčních pravidel: regular grammars, logic rules

EXTRAKČNÍ ONTOLOGIE

Struktura

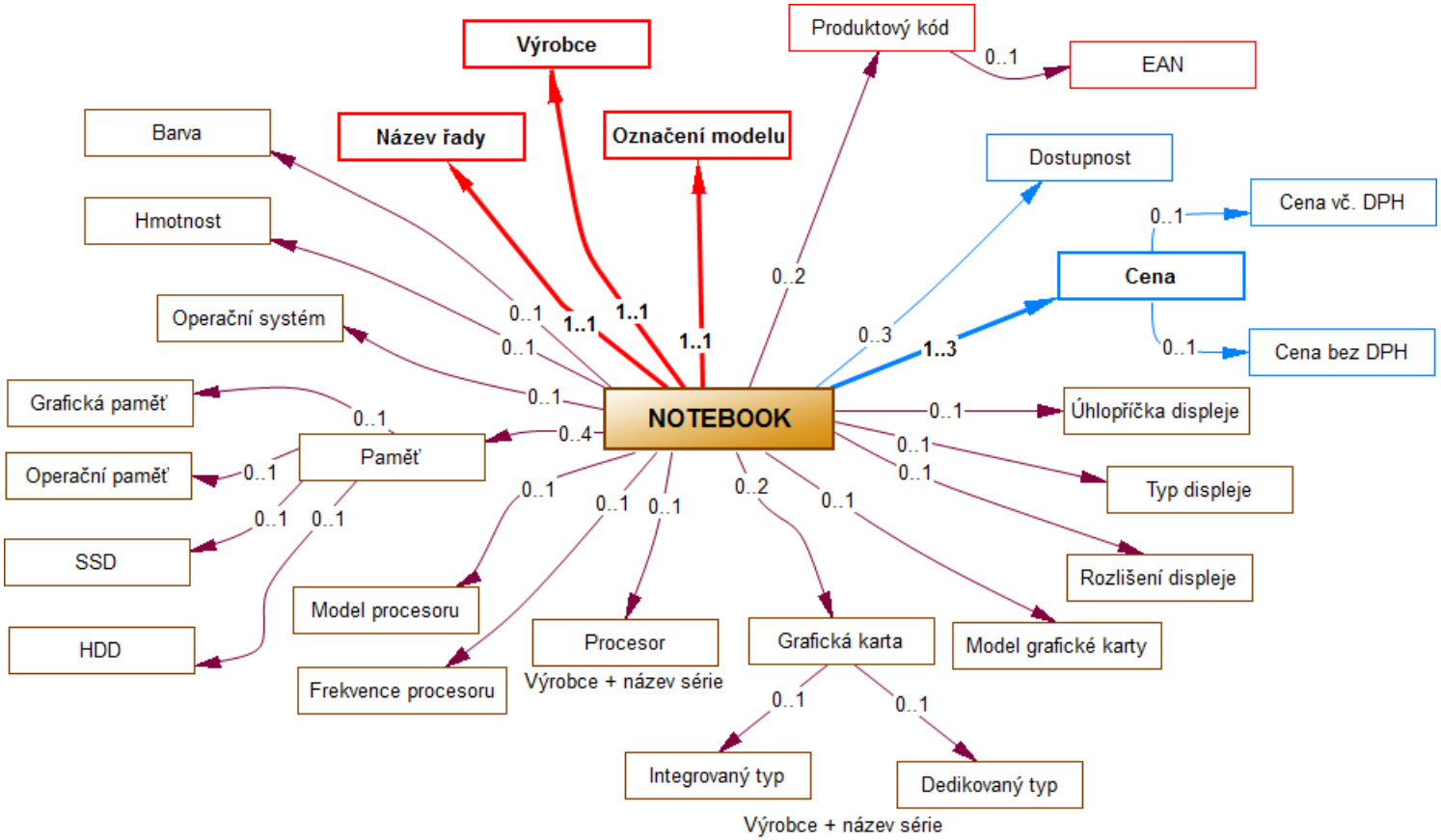
- XML schéma
- definice tříd, atributů, samostatných atributů
- specializace atributů

Extrakční evidence

- expert: ručně zadaná pravidla (znalosti o doméně)
- učící algoritmy: trénované modely (trénovací data)
- systém: indukce wrapperů (DOM strom dokumentu)
    - pravidelná formátovací struktura

EXTRAKČNÍ ONTOLOGIE

Doména produktových nabídek / koncept Notebook



RUČNĚ ZADANÁ PRAVIDLA

Pravidla (vzory)

- definice vzoru: přesnost (precision) a pokrytí (cover)
  - positivní vs. negativní vzory
- pravidla na úrovni třídy
  - seskupování atributů do instancí třídy
- pravidla na úrovni atributů
  - dva typy: hodnotové a kontextové vzory
  - literály, regulární výrazy, formátovací prvky (HTML tagy) a formátovací vzory, axiomy (JavaScript)
  - min./max. hodnoty (u čísel)
  - min./max. počet slov (u textových řetězců)
  - koreferenční rozhodnutí
  - rozsáhlé seznamy slov (gazetteery)

RUČNĚ ZADANÁ PRAVIDLA

Příklad pravidla (pro atribut *název pracovní pozice*):

Kontextový vzor (pravidlo):

```
<pattern id=„ukazka“ p="0.4" cover="0.4">
  ( <lab name="^HEADING|TITLE|LI|TD|STYLE|A|B|EM|SPAN|STRONG"/>
  <tok/>{0,2} $ <tok/>{0,2}
  <lab name="HEADING|TITLE|LI|TD|STYLE|A|B|EM|SPAN|STRONG$"/> )
  ( pozici | pozice ) ( <tok/>{0,2} $position ){0,8} <tok/>{0,2} $
</pattern>
```

Vzor obsahuje dva „vnořené“ vzory, z nichž musí být splněn alespoň jeden. Znak dolaru ($) je zástupný znak pro kandidáta na hodnotu atributu. Kontextový vzor tedy bude splněn, pokud:

- kandidát na hodnotu ($) se nachází uvnitř jednoho z přípustných HTML tagů (nalevo a napravo od kandidáta jsou nanejvýš dva libovolné tokeny) NEBO
- před kandidátem je buď jedno z přípustných slov (pozici, pozice) anebo je kandidát součástí posloupnosti (nanejvýš osmi) již extrahovaných názvů pozic (označených jako $position a oddělených nanejvýš dvěma tokeny), přičemž před první extrahovanou hodnotou v této posloupnosti je jedno z uvedených slov (pozici/pozice).
- Vyhovuje všem 3 názvům pozic uvedeným ve větě: *„Hledáme uchazeče na pozice číšník, kuchař nebo šéfkuchař“.*

KLÍČOVÉ ÚLOHY

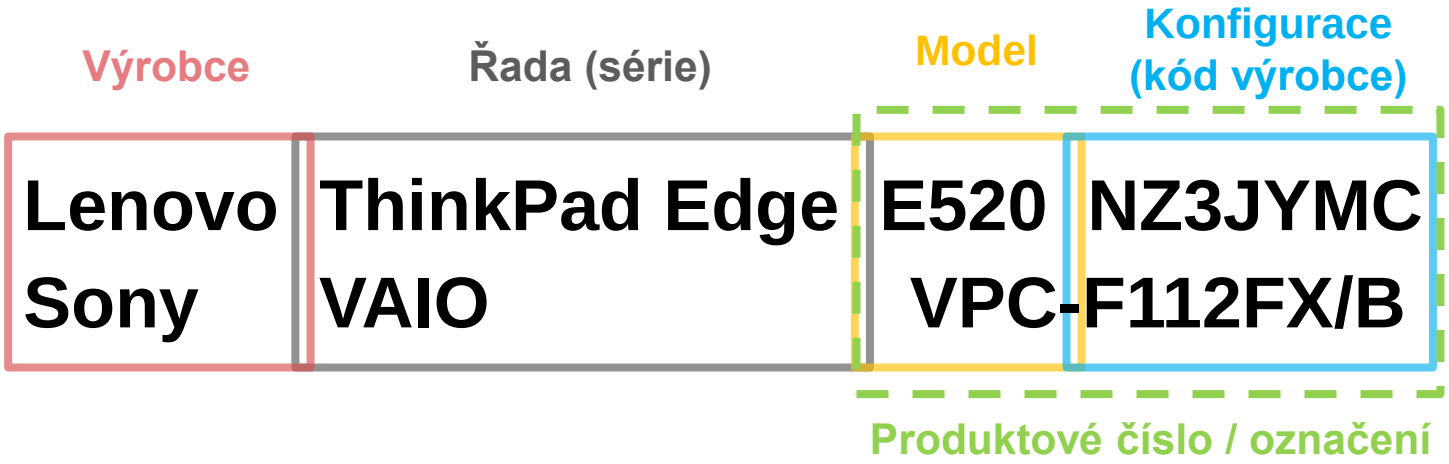
Klíčové úlohy

- Rozpoznání relevantního dokumentu
- Rozpoznání obsahové části dokumentu
- Vytvoření pravidel – přesnost vs. úplnost
- Vytvoření potřebných slovníků
- Seskupení atributů do instancí

VARIACE ÚDAJŮ V NÁZVECH PRODUKTŮ

Různé systémy značení produktů napříč výrobci

- mění se s časem – prolínání starých s novými



E-shopy vytvářejí další varianty

- vynecháním údaje
- přehozením dvou údajů
- vložením údaje nesouvisejícího s názvem (barva)
  Lenovo ThinkPad Edge E520 červený 1143-JYG

HEUREKA.CZ – JEDNOTNÁ PRAVIDLA

Sekce	Povinné údaje
Autosedačky	Výrobce | Řada | Produktové číslo / Označení | Barva | Ročník
Notebooky	Výrobce | Řada | Produktové číslo / Označení
PC komponenty	Výrobce | Řada | Produktové číslo / Označení | Vzor
Baterie	Výrobce | Řada | Produktové číslo / Označení | Množství | Určení / Typ
Satelitní komplety	Výrobce | Řada | Verze
Jízdní kola	Výrobce | Řada | Produktové číslo / Označení | Barva | Rozměry / Velikost | Ročník
Káva	Výrobce | Množství

Zdroj: Heureka.cz

GRANULARITA ÚDAJŮ

- **product matching**
  - identifikace produktů a spárování
  - globální produktový kód (EAN)
    - je k dispozici?
  - čím více informací, tím robustnější
    - např. překlep v EAN kódu nebo nějakém parametru
  - gazetteer list nebo obecný vzor?
    - jak často se hodnoty atributu mění?
    - lze zachytit regulárním výrazem? (název řady vs. číslo modelu)
    - obor hodnot

NÁZVY PRACOVNÍCH POZIC

Charakteristika

- otevřený obor hodnot („nekonečně“ mnoho kombinací názvů povolání a podpůrných slov)
- libovolná délka (jednoslovné i desetislovné)
- průměrná délka pracovní pozice na jobs.cz: 4,17 slov (vzorek 13 000 záznamů)

Pracovní vymezení pojmů

- profese: druh pracovní činnosti, tvořený jediným slovem
  - *architekt, ekonom, inženýr, právník, pracovník, …*
- pracovní pozice: rozvinutí názvu profese o další slova nebo sloučení dvou profesí dohromady
  - *daňový poradce, bankovní poradce, specialista pro oceňování vozidel, inženýr ekonom, …*

NÁZVY PRACOVNÍCH POZIC

Složený název pracovní pozice

- více profesí oddělených interpunkčním znaménkem (lomítko, čárka, pomlčka)
  - *asistentka vedení společnosti/asistentka jednatele*
  - *tlumočnice/asistentka*
- význam „oddělovače“ – složený název jedné pozice nebo posloupnost více pozic?
  - *technolog, průmyslový inženýr*
  - *hledáme číšníky, kuchaře, šéfkuchaře*
- volba mezi přesností a úplností
- možná řešení
  - slovník (tezaurus, ontologie) zachycující relace mezi jednotlivými typy povolání
  - příp. „jemné“ seskupování profesí podle podobnosti náplně práce)

EXTRAKCE NÁZVŮ POZIC

1) **„volná“ extrakce**
   - extrakce celé sekvence slov mezi takovými dvěma HTML tagy, které „zvýrazňují“ textový obsah (např. HEADING), obsahuje-li tato sekvence *název profese*
   - možnost podmínit extrakci jen při určitém počtu slov mezi tagy
   - čím „významnější“ element, tím vyšší přiřazená váha
   - výhoda: není potřeba anotovat, stačí slovník profesí
2) **striktní extrakce**
   - slovník pokrývající všechny možné varianty názvů pozic
   - hledání začátku a konce pracovní pozice
     - podle positivního slovníku
     - algoritmy strojového učení – LP^2, CRF, SVM
     - nevýhoda: dostatek trénovacích dat
3) **kombinace obou přístupů**

UKÁZKA VÝSTUPU

„volná“ extrakce dle okolních HTML tagů

- Vstup: 10 107 dokumentů.
- Výstup: 16 665 nalezených pozic v 8579 dokumentech.

BI vývojář	www.neit.cz	0.9950	65
Senior tester	www.neit.cz	0.9950	65
Programátor PL / SQL	www.neit.cz	0.9950	65
Administrátor	www.neit.cz	0.9950	65
Junior tester	www.neit.cz	0.9950	65
pečovatel / ka .	www.psrakovnik.cz	0.9950	66
Mistr ve výrobě (Group Leader)	www.kmcz.cz	0.9950	68
Kvalifikovaný operátor - obraběč kovů , brusič	www.kmcz.cz	0.9950	68
Vedoucí údržby (strojař)	www.kmcz.cz	0.9950	68
Kvalifikovaný operátor - lakovna	www.kmcz.cz	0.9950	68
Projektant - Elektro	www.meritumkladno.cz	0.9950	69
Finanční specialista RSTS	www.rsts.cz	0.9950	70
ANALYTIK / ANALYTIČKA ÚVĚROVÉHO RIZIKA	www.rsts.cz	0.9950	70
SPECIALISTA / SPECIALISTKA PODPORY PRODEJE	www.rsts.cz	0.9950	70
TRENÉR / TRENÉRKA	www.rsts.cz	0.9950	70
Sales manager	www.inzerce-ceskobudejovicko.cz	0.9950	71
Recepční do hotelu na Praze 1	www.inzerce-ceskobudejovicko.cz	0.9950	71

SESKUPOVÁNÍ ATRIBUTŮ DO INSTANCÍ

Přístup „bottom-up“

- pravidla na úrovni třídy definují možné permutace atributu (X hledání datových záznamů na základě analýzy fyzické struktury)

Pracovní nabídky

- v případě více pracovních nabídek na stránce s odlišnými permutacemi atributů obtížné nalézt hranici
- čím více pravidel, tím menší rozlišovací schopnost
- u strukturovaných nabídek členěných do logických celků lze využít klíčových slov jako *pracovní nápň, požadavky na uchazeče, nabídka*

UKÁZKA

Job 1641 (1,0000/0,0000)

position	STROJNÍ KONSTRUKTÉR TECHNOLOG	(0.994975174)
tmp_word	Naše požadavky na pracovníka :	(1)
contract	HPP	(0.989479321)
orphans: [25,43] Path score=-0,214		
praxis_in_yrs	5	(1)
language	Aj	(0.588855365)
orphans: [51,66] Path score=-0,214		
language	Nj	(0.588855365)
drivingLicense	B	(0.912663855)
praxis	nutná	(0.996468403)
tmp_word	Popis pracovní náplně :	(1)
starting_day	co nejdříve	(0.996468503)
null		
null		

Job 1659 (1,0000/0,0000)

position	STROJAŘ ZÁMEČNÍK	(0.994975074)
tmp_word	Naše požadavky na pracovníka :	(1)
contract	HPP	(0.989479321)
orphans: [158,176] Path score=-0,214		
praxis_in_yrs	5	(1)
language	Aj	(0.588855365)
orphans: [184,199] Path score=-0,214		
language	Nj	(0.588855365)
drivingLicense	B	(0.912663855)
praxis	nutná	(0.996468403)
tmp_word	Popis pracovní náplně :	(1)
starting_day	co nejdříve	(0.996468503)
locality	Nový Bydžov	(0.947506762)
null		

STROJNÍ KONSTRUKTÉR TECHNOLOG

Naše požadavky na pracovníka:

-vzdělání:SS, VŠ- strojírenství (není podmínkou) ale je to lepší
-práce na HPP
-praxe:min. 5 let v oboru (není podmínkou) ale čím více praxe tím lépe
-jazyky:Aj nebo Nj
-certifikáty či průkazy – ŘP sk. B

ostatní požadavky:
-Znalost 2D/3D CAD systému, Solidworks
-praxe a vzdělání v oboru nutná

Popis pracovní náplně:
Nabízíte:
-Samostatnou práci
-Dobré platové podmínky

- Nástup možný co nejdříve
- Pracoviště Červeněves

Mzdové rozpětí a benefity:
Měsíční nástupní plat 18-35 000 tisíc Kč- dle praxe a zkušeností
Různé zaměst. výhody: Stravenky , příspěvky na spoření ING, věrnostní program

---

STROJAŘ – ZÁMEČNÍK

Naše požadavky na pracovníka:

-vzdělání:USO, SS- strojírenství (není podmínkou) ale je to lepší
-práce na HPP
-praxe:min. 5 let v oboru (není podmínkou) ale čím více praxe tím lépe
-jazyky:Aj nebo Nj vítáno
-certifikáty či průkazy – ŘP sk. B

DÍLČÍ EXPERIMENTY

Vytváření slovníku profesí

- získání cca 400 jednoslovných názvů profesí z katalogu NSP
- automatické vygenerování názvů profesí v ženském rodě
- další možnosti vylepšení:
  - obohacení slovníku o další slovní tvary profesí (X morfologická desambiguace)

Analýza názvů pracovních pozic & jejich okolí

- alternativa k učícím algoritmům ML
- hledání typických „začátků“ a „konců“ pozic
  - výhoda: není třeba anotovat - možnost využít data získaná DOM parsingem z jobs.cz (13000 záznamů)
- hledání typického okolí pozice (první slovo „před“ a „za“)
  - nevýhoda: vyžaduje opět anotovaná data

SHRNUTÍ

Výhody

- kombinací regulárních výrazů a formátovacích prvků je možné extrahovat entity s otevřeným oborem hodnot
- možnost kombinovat ručně zadaná pravidla s algoritmy ML
- lze vytvořit rychlý a funkční prototyp extrakční ontologie i bez trénovacích dat

Nevýhody

- obtížné seskupit atributy do instancí, jejichž pořadí je libovolné (v případě produktových katalogů i pracovních nabídek)

Možnosti vylepšení

- analýza fyzické struktury dokumentu (hledání opakující se struktury) – možnost využít u automaticky generovaných záznamů podle šablon (produktový katalog)
- algoritmus pro indukci wrapperů – viz následující slide

OBCHODNÍ ZÁSTUPCE pro prodej komodit

Požadujeme:
- min.středoškolské vzdělání, ŽL
- komunikativnost
- znalost základní práce s PC
- veselou, optimistickou povahu
- slušné vystupování a důvěryhodný vzhled
- učenlivost

Nabízíme:
- perspektivní práci v dynamicky se rozvíjejícím oboru
- možnost karierního růstu
- zajímavé odměňování, na základě individuálních výsledků
- týmové odměny
- osobní vzdělávání ve firemním vzdělávacím kursu

Spolupracovník FINAPS Consulting

Požadujeme:
- důvěryhodnost
- touhu vylepšit si svoji finanční situaci

Nabízíme:
- zajímavou finanční odměnu

Finanční poradce

Požadujeme:
- min. středoškolské vzdělání,ŽL + ŘP sk. B
- komunikativnost a příjemné vystupování
- zodpovědný osobní přístup a kreativitu při plnění úkolů
- samostatnost a vysoké osobní nasazení
- základní orientace na kapitálových trzích výhodou

Nabízíme:
- zajímavou práci ve stále se měnícím investičním prostředí
- zajímavé ohodnocení odpovídající osobnímu nasazení
- vyškolení,vzdělávání a stálý servis po dobu trvání smluvního vztahu
- flexibilní pracovní dobu
- možnost karierního postupu

Indukce wrapperů: 1 z 5 pozic nebyla rozpoznána ani přes shodnou formátovací strukturu v okolí hodnot.

DĚKUJI ZA POZORNOST