Detekce vizuálních vzorů ve webových stránkách

Autor: Ing. Martin Kotraš
Vedoucí: doc. Ing. Radek Burget, Ph.D.

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Motivace

- Webové stránky jsou obrovský zdroj informací.
- Strojové zpracování informací z nich v současné době vyžaduje ruční vytvoření speciálních programů (scraperů) pro každý jednotlivý zdroj.
- To je pracné, obtížně škálovatelné a chybové.

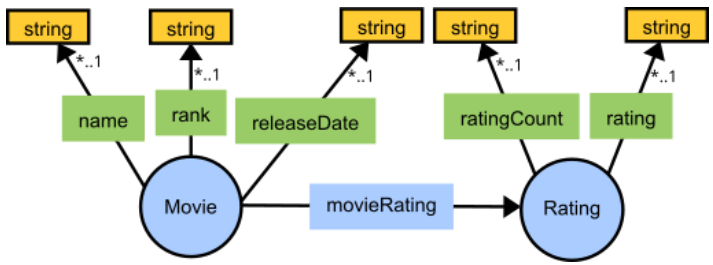
Cíle práce

- Automatická identifikace důležitých údajů na libovolné webové stránce.
- Využití obecného modelu očekávané informace.
- Detekce opakujících se vizuálních vzorů na webové stránce, které mohou reprezentovat očekávanou informaci z obecného modelu.

Vstup aplikace

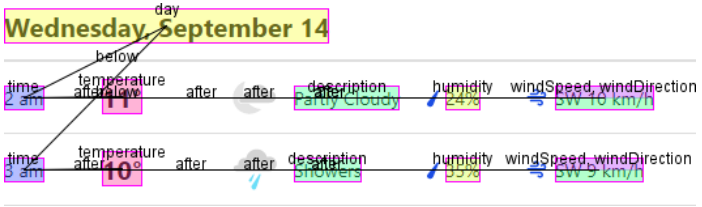
- URL adresa webové stránky k extrakci.
- Ontologický datový model hledaných informací.

```
./vizget https://www.csfd.cz/zebricky/ \
  < movie-ontology.ttl > extracted.ttl
```



Výstup aplikace

- Extrahované informace navázané na dat. model.



Postup extrakce

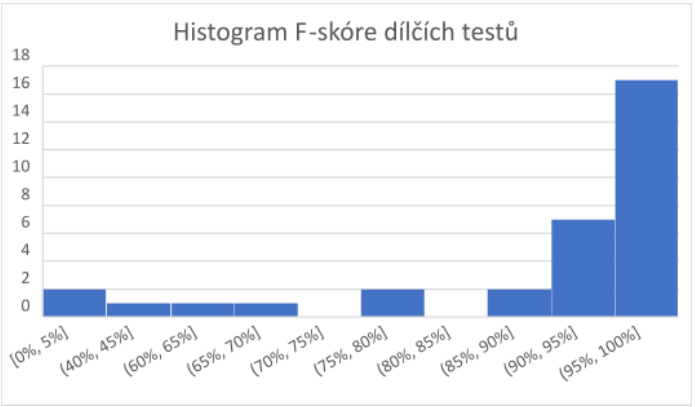
1. **Vykreslení dokumentu** vykreslovacím jádrem.
2. **Segmentace oblastí** – převod ohraničujících boxů prvků stránky na oblasti o jednom a více boxech; spojení víceřádkových oblastí.
3. **Označování oblastí** značkovači z dat. modelu (regulární výraz, slovník slov, datum, osoby, …).
4. **Vyhledání vzorů** v označených oblastech.
5. **Generování výstupu** – RDF, obrázek.

Algoritmus vyhledávání vzorů

- **Vstup:** označené oblasti, dvojice vlastností.
- **Výstup:** množina vzorů, tj. dvojic oblastí s prvky stejného vizuálního stylu a prostorového vztahu.
- Výběr nejlepší kombinace stylů a prostorových vztahů metrikami: počet vzorů v množině, unikátnosti textu identifikující oblasti, …
- Odstranění nežádoucích vzorů reduktory: založené na kardinalitě vlastností, vzdálenosti mezi prvky na stránce, ...

Výsledky experimentů

- Aplikace byla testována celkem na **7** doménách: žebříček filmů, novinky z fakultního webu, předměty fakulty, e-shop, jízdní řád, předpověď počasí a program rozhlasových stanic.
- Celkem bylo vytvořeno **33** dílčích testů:
  - **~45 %** z nich uspělo s F-skóre **100 %**,
  - **~75 %** z nich získalo nad **85 %** F-skóre,
  - **~90 %** z nich získalo nad **60 %** F-skóre.



Přínosy práce

- Vznikla prakticky použitelná aplikace pro automatickou extrakci informací z webu.
- Vylepšen byl současný algoritmus vyhledávání vzorů o hodnocení metrikami a odstranění nežádoucích vizuálních vzorů reduktory.
- Práce může sloužit jako základ pro vědeckou publikaci a další výzkum v oblasti extrakce.