Towards Web Semantization and User Understanding

Ladislav Peška, Ivo Lašek, Alan Eckhardt, Jan Dědek, Peter Vojtáš and Dominik Fišer

Department of Software Engineering, Charles University in Prague, Czech Republic
Outline

● Motivation
● Web Semantization – Our Approach
● Extractors and Annotators
● User Agents
● Conclusion and Future Work
Motivation – Semantic Web

- Idea from Berners-Lee, Hendler, Lassila, 2001
- Semantic content made by web creators
  - Not working well
  - Several changes throughout the time
    - However there is still need for machine readable web content
    - Today rather process of semantization of the web
    - Made by community
    - Linked data
Web Semantization – Our First Approach

- Generic web crawler, crawl whole czech web
- Various semantic extractors
- No user aspect
- No intended purpose of the data
- Who creates ontologies?
Web Semantization – Our Vision
Web Semantization – Our Vision
Web Semantization – Our Vision
Web Semantization – Our Vision
Web Semantization – Our Vision
Web Semantization – Our Vision

- Employ users
  - As a source of some semantical data
  - As consumers of added value
- Semantic data should have reason / application model why to be collected
- Several tools processing parts of our model developed, initial integration steps
Semantic Data Extracting

- Tools to gather unstructured and semi-structured data:
  - Information Extraction from Natural Language
  - Information Extraction based on structural similarity
  - Domain dependant annotations
Semi-Structured Data Extracting

- To identify objects on e.g. e-commerce category page:
  - Contains more objects (records) of the same type
  - The records have similar structure (DOM)
  - The attributes can be identified via ontology + RegExp
  - Record similarity via Levenshtein distance

- Several problems occurred e.g. records are not trees, but forests
Semi-Structured Data

Extracting - Example
Unstructured Information Extracting

- From natural language – e.g. news articles
  - Several linguistic tools (tokenizer, morphological analyzers…)
  - Named entity recognition, data aggregation, new attributes etc.

Example of tectogrammatical linguistic tree. Source sentence: “The new 2011 Audi A8 has an Bang & Olufsen Advanced sound system with 19 speakers, 19 channels and more than 1400 watts.”

Named entities Audi A8 and Bang & Olufsen are decorated by green color over corresponding nodes.
Unstructured Information Extracting – News Recommending

- Outsourcing named entities

Goldman Sachs Rises as Investors Bet on Comeback

Goldman Sachs Group Inc. (GS) rose 5.5 percent in New York trading as investors looked past a third-quarter loss and focused on gains in trading revenue and prospects for a rebound in underwriting and takeovers.

dbpedia-owl:industry dbpedia:Financial_services
dbpprop:locationCity New York City
fb:organization.organization.date_founded 1869
News Recommending - User Profile Creation

- Use named entities and their properties as features of an article
- Various machine learning approaches to learn user interests
  - Association rules mining
  - Clustering
- Formal Concept Analysis
Domain Specific User Annotations

- Annotations based on ontology specified by user
- Collaborative benefit from other users annotations
- Work in progress on machine annotating of similar pages
User Agents - Overview

- Searching & Recommending
- Data from either an IE method or obtained from site directly
- Other semantic data (e.g. Named Entities) possible
E-Commerce Recommending

- Motivation from e-shops
  - Recommendations are beneficial for both user and site owner
  - However it is often too complicated to employ recommending

- Develop simple to use recommending component
  - Deployable into the running e-commerce sites
  - Independent on e-shop domain, object attributes etc.
  - Extendible into the multi-site recommender
E-commerce Recommending – Explicit Preference Model
E-commerce Recommending – Explicit Preference Model

- Decompose user preference (user rating)
  - Attribute preference, importance of the attributes
- Learn how to aggregate attribute preferences
E-Commerce Recommending – Implicit Feedback Factors

- Recently added into our preference model
  - User Rating is not numerous enough
  - We can record also other (implicit) user feedback factors
    - Time on page,
    - Mouse clicks, mouseover etc.
    - Clickstream,
    - Scrolling,
    - Recommending related actions
    - Purchase related actions (buy, add to chart) etc.
- How to use and combine them?
E-commerce recommending – implicit feedback factors

- Similar idea as for attribute preferences
- Learning from positive examples (purchases)
E-commerce recommending – feedback factor experiments

UserPreference = \( \frac{9 \times \text{MouseClicksPref}() + 5 \times \text{PageViewsPref}() + 4 \times \text{TimeOnObjectPref}()}{18} \)
Conclusions & Future work

- Idea of web semantization
  - We have both methods for semantic IE, use-cases how to use them and user agents implementing them
    - Implemented, tested, experimental data

- The links between tools are mostly not finished

- Work to be done on:
  - Annotator (machine annotations of similar pages)
  - E-commerce recommender (implementing feedback factors preference model)