

Web Information Extraction Systems for Web Semantization

Jan Dědek^{1,2}

¹Department of Software Engineering, Faculty of Mathematics and Physics,
Charles University in Prague, Czech Republic

²Institute of Computer Science, Academy of Sciences of the Czech Republic

ITAT 2009

28 September, Kralova studna

Outline

1 Introduction

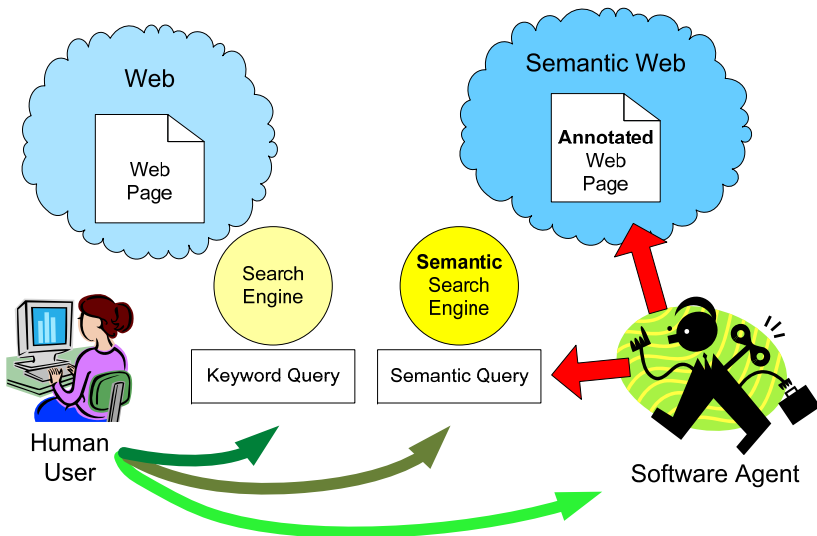
- The Semantic Web in Use
- Web Semantization

2 Web Information Extraction

- Web Information Extraction Approaches
- User Initiative and Effort
- Information Extraction based on Web Page Structure
- Information Extraction from Text-based Resources

3 Conclusion and Future Work

The Semantic/Semantized Web in Use

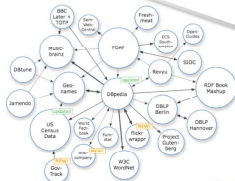


Growing of Linking Open Data data set 2007–2008

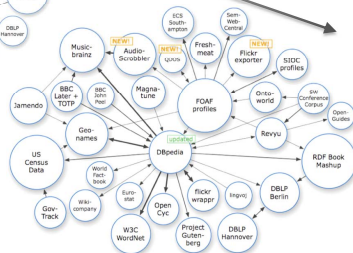
2007



Motivation



2008



M. Hausenblas

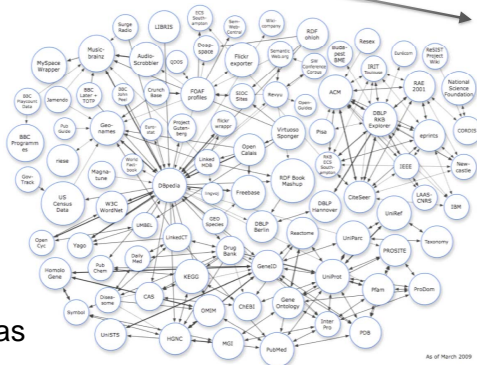
Growing of LOD data set 2008–2009

2008



Motivation

2009



M. Hausenblas

As of March 2009

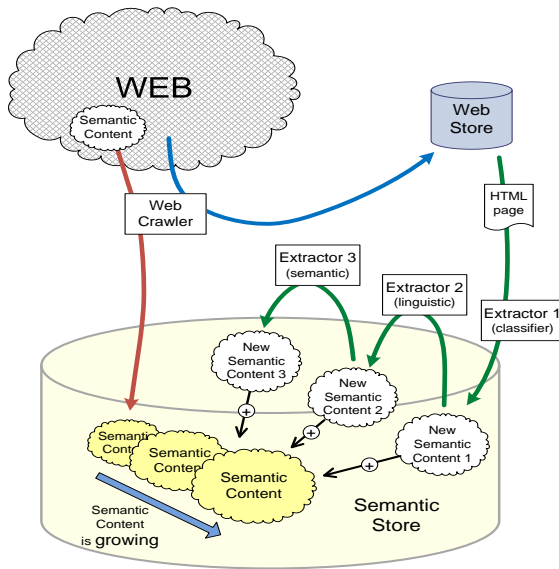
LOD data set statistics as of July 2009



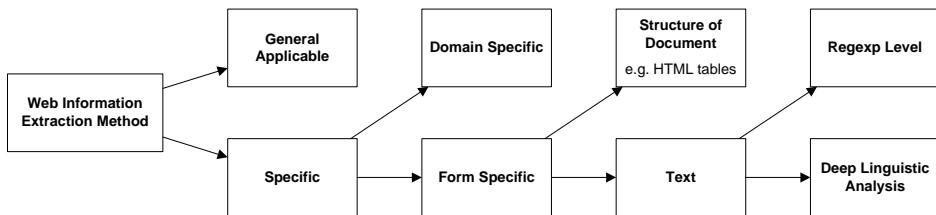
<i>Domain</i>	<i>No of Triples</i>	<i>% of Cloud</i>	<i>No of Links</i>	<i>% of Links</i>
Media	698.000.000	10,4%	1.238.000	0,8%
Publications	212.000.000	3,2%	4.922.000	3,3%
Life Sciences	2.429.000.000	36,1%	133.199.000	89,4%
Geographic Data	3.097.000.000	46,0%	4.038.000	2,7%
User Generate Content	76.000.000	1,1%	1.559.000	1,0%
Cross-Domain	214.000.000	3,2%	3.992.000	2,7%
Total	6.726.000.000		148.948.000	

Christian Bizer: The Web of Linked Data (26/07/2009)

Web Semantization

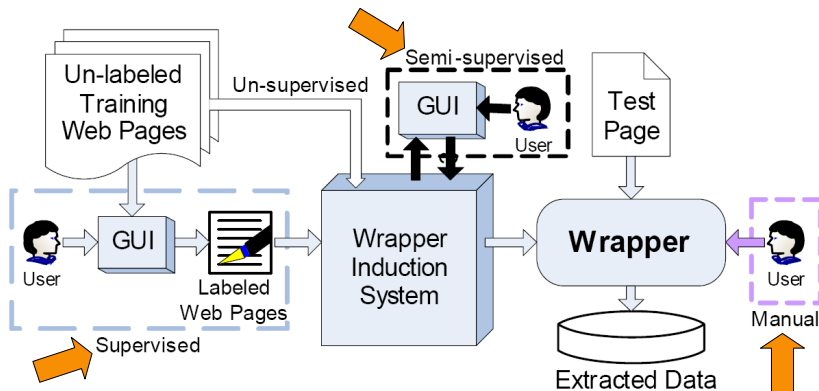


Division of extraction methods



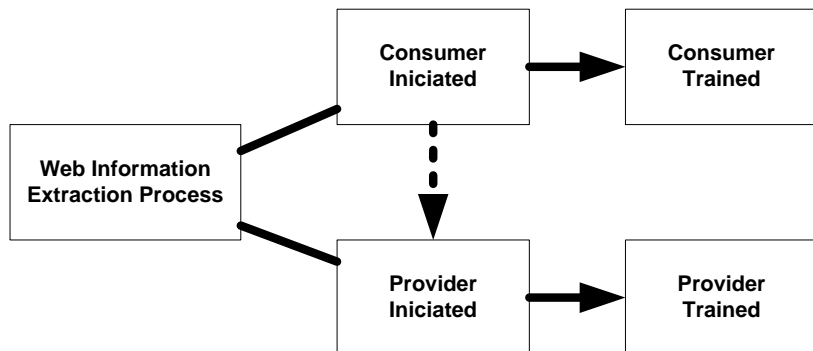
- General Applicable
 - Instance Resolution Task
 - Bootstrapping
 - Use of search engines
- Domain Specific
- Form Specific

A general view of WI systems – user perspective



Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, Khaled F. Shaalan,
"A Survey of Web Information Extraction Systems," IEEE Transactions on
Knowledge and Data Engineering, vol. 18, no. 10, pp. 1411-1428, October, 2006.

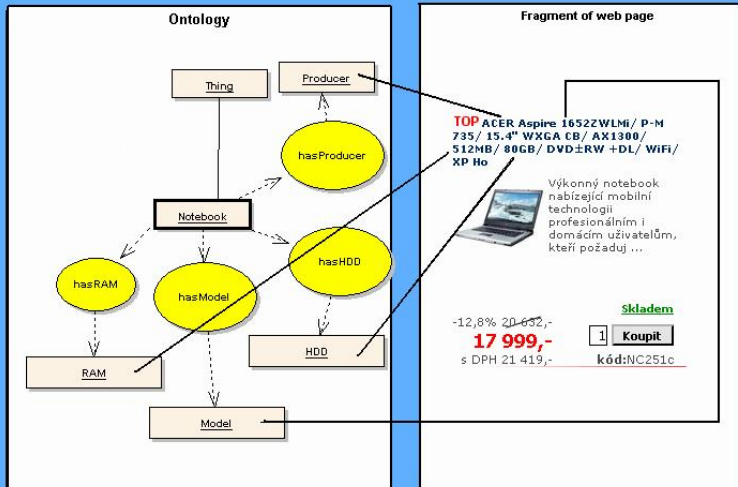
User initiative and effort – Web Semantization



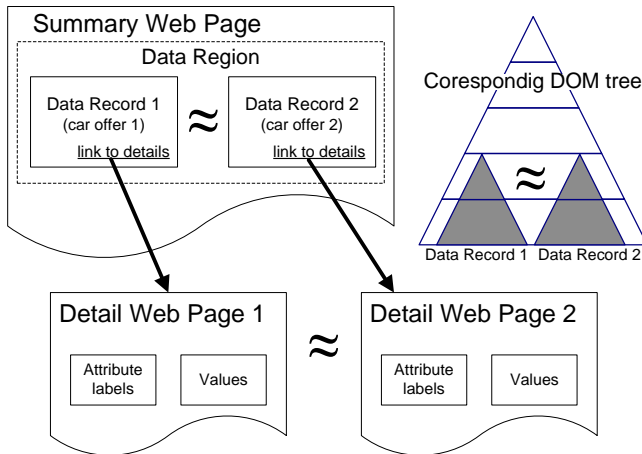
- Not only Information Extraction
- But also Semantic Annotation
- Domain specific knowledge has to be **obtained** in all cases.
- Scalability?

Information Extraction based on Web Page Structure

Extraction Based on Structural Similarity (suitable for Web Shops)



Extraction Based on Structural Similarity – extraction method



Benchmarks for Information Extraction

- No common benchmark for “structured pages” or Web IE.
- Message Understanding Conference (MUC)
- Automatic Content Extraction (ACE) Evaluation
- Text Analysis Conference (TAC)
- Text REtrieval Conference (TREC)
- Document Understanding Conferences
(text summarization)

Classical tasks of text preprocessing and linguistic analysis

Text Extraction – e.g from HTML, PDF or DOC,

Tokenization – detection of words, spaces, punctuations, etc.,

Segmentation – sentence and paragraph detection,

POS Tagging – part of speech assignment, often including
lemmatization and morphological analysis,

Syntactic Analysis (often called linguistic *parsing*) –
assignment of the grammatical structure to given
sentence with respect to given linguistic formalism
(e.g. formal grammar),

Coreference Resolution (or *anaphora resolution*) – resolving
what a pronoun, or a noun phrase refers to. These
references often cross boundaries of a single
sentence.

Classical domain dependent IE tasks

Named Entity Recognition: This task recognizes and classifies named entities such as **persons**, **locations**, **time expression**, or **measuring units**.

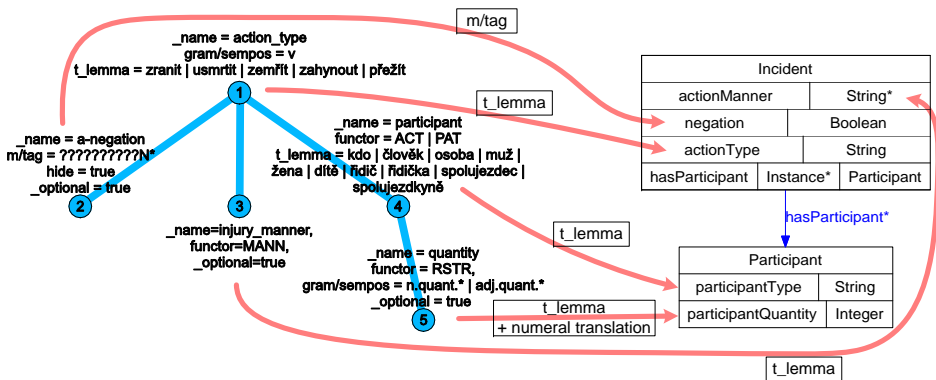
Template Element Construction: Populates templates describing entities with extracted **roles** (or **attributes**) about one single entity.

Template Relation Construction: As each template describes information about one single entity, this task identifies semantic **relations between entities**.

Template Unification: **Merges** multiple elementary templates filled with **information about identical entities**.

Scenario Template Production: Fits Template Elements and Template Relations into templates describing pre-specified event scenarios (**pre-specified “queries on the extracted data”**).

Linguistic IE and Semantic interpretation of extraction rules



- Determines how particular values of attributes are used.
- Gives semantics to extraction rule.
- Gives semantics to extracted data.

Linguistic IE – ILP Learning of Extraction Rules

Zprávy Jihočeského kraje

Zubatelno 1, 614 00 Brno, telefon 950 630 111,
<http://www.zpravy.cz>
 Zpravodajství v roce 2006

15.05.2007

V trabantu zemřeli dva lidé
 K tragické nehodě dnes odpoledne hasiči vyjžděli na silnici z obce Česká Krumlov na Brněnsku.

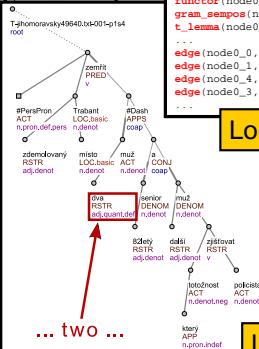
Nehoda byla operativně středisku HZS ohlášena ve 13.13 hodin. Na místě zasahovala jednotka profesionálních hasičů ze stanice Hájovské. Jednalo se o černí sračku autobusu Karosa s vozidlem Trabant. Podle dostupných informací trabant jedoucí ve z Elna do Krumlova vyjel do protisměru, kde narazil do liniového autobusu dopravní společnosti ze Žďaru nad Sázavou. Ve zdemolovaném trabantu i místě zemřeli dva muži – 82letý senior a další muž, jehož totožnost určili policisté.

Hasiči odstraňovali vozidla přírodním způsobem a po vyčištění zdemolovaného autobusu dopravní policie vrak trabantu zaklesnutý po autobusem pomocí lana odtrhla. Po odstranění střechy trabantu pak hasiči vyprostoili těla obou mužů. Obě vozidla – trabant i autobus, po postupné odstranění na kraj vozovky, a uvolnění tak jeden jízdní pruh. U liniových kapalin nebyl zjištěn. Po 16. hodině pomohli vrak trabantu přeložit k odstavu a asistovali při odtažení autobusu. Po ukončení vozovky bylo před 16.30 hod. místo nehody předat policistům a uvolnění vozovky.

Source web page

```
tree_root(node0_0). node(node0_0).
id(node0_0, t_jihomoravsky49640_txt_001_pls4).
***** node0_1 *****
node(node0_1).
functor(node0_1, pred).
gram_sempos(node0_1, v).
t_lemma(node0_1, zemrit).
***** node0_2 *****
node(node0_2).
functor(node0_2, act).
gram_sempos(node0_2, n_pron_def_pers).
t_lemma(node0_2, x_perspron).
***** node0_3 *****
node(node0_3). id(node0_3,
functor(node0_3, loc).
gram_sempos(node0_3, n_denot).
t_lemma(node0_3, trabant).
...
edge(node0_0, node0_1). edge(node0_1, node0_2).
edge(node0_1, node0_3). edge(node0_3, node0_4).
edge(node0_4, node0_5). edge(node0_3, node0_6).
edge(node0_3, node0_7). edge(node0_3, node0_8).
```

Logic representation



... two ...

Linguistic trees

Conclusion and Future Work

Conclusion:

- Partial survey of WIE systems
(see the paper for references)
 - Related to **Web Semantization**
- Problem of **unskilled user** pointed out

Future Work:

- Future **development of WIE** tools and work on their adaptability to new domains.
- **Integration** of WIE tools to the web semantization system.
- Development of the methodology and software to support the **extension** of the semantization system **to a new domain for a non-expert user**.