

---

# Content based user centric model of a network search system

---

Peter Vojtáš

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

# Outline

---

- Motivation examples
- The Idea of Web Semantization
  - Semantic Web and Web Semantization
  - **Domain independent intermediate** annotation followed by **domain dependent** annotation
  - Proposed architecture of the system, repositories
- Methods of Semantic Annotation
- User Preferences on the (Semantic) Web
- Department of Software Engineering overview

# Motivation example – buying a notebook

---

**User** (me or maybe my sister) wants to buy a notebook.

**Requirements** - **either known** - CPU, memory, HDD, display, size, weight, manufacturer, color, price, etc.  
- **or unknown**

**Also interested** in used notebooks and other users comments and reviews.

Our Semantic **repository** contains semantically Annotated information from **several** web shops, online auctions, user textual comments (sentiment analysis), blogs, etc.

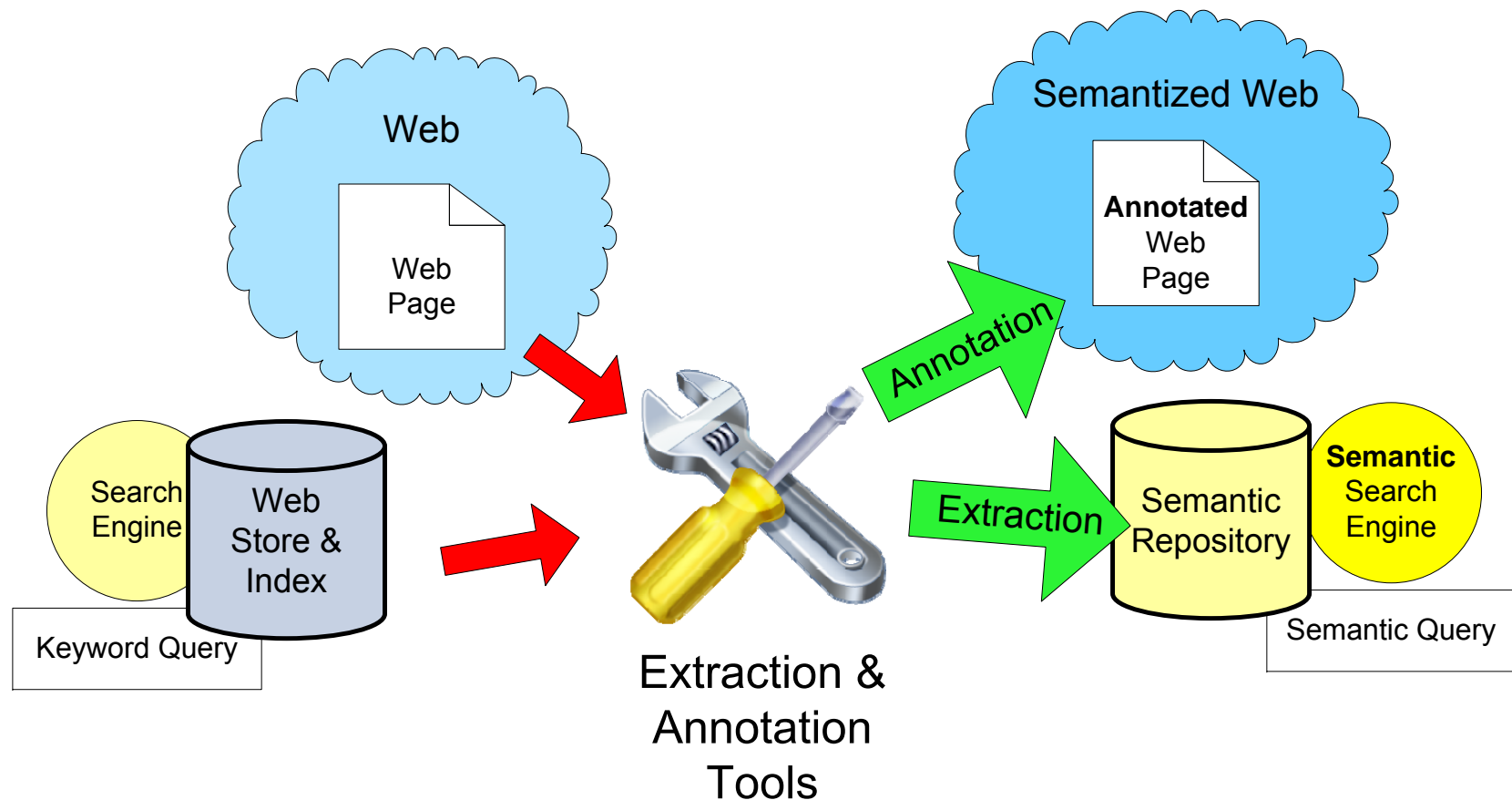
# The Idea of Web Semantization

---

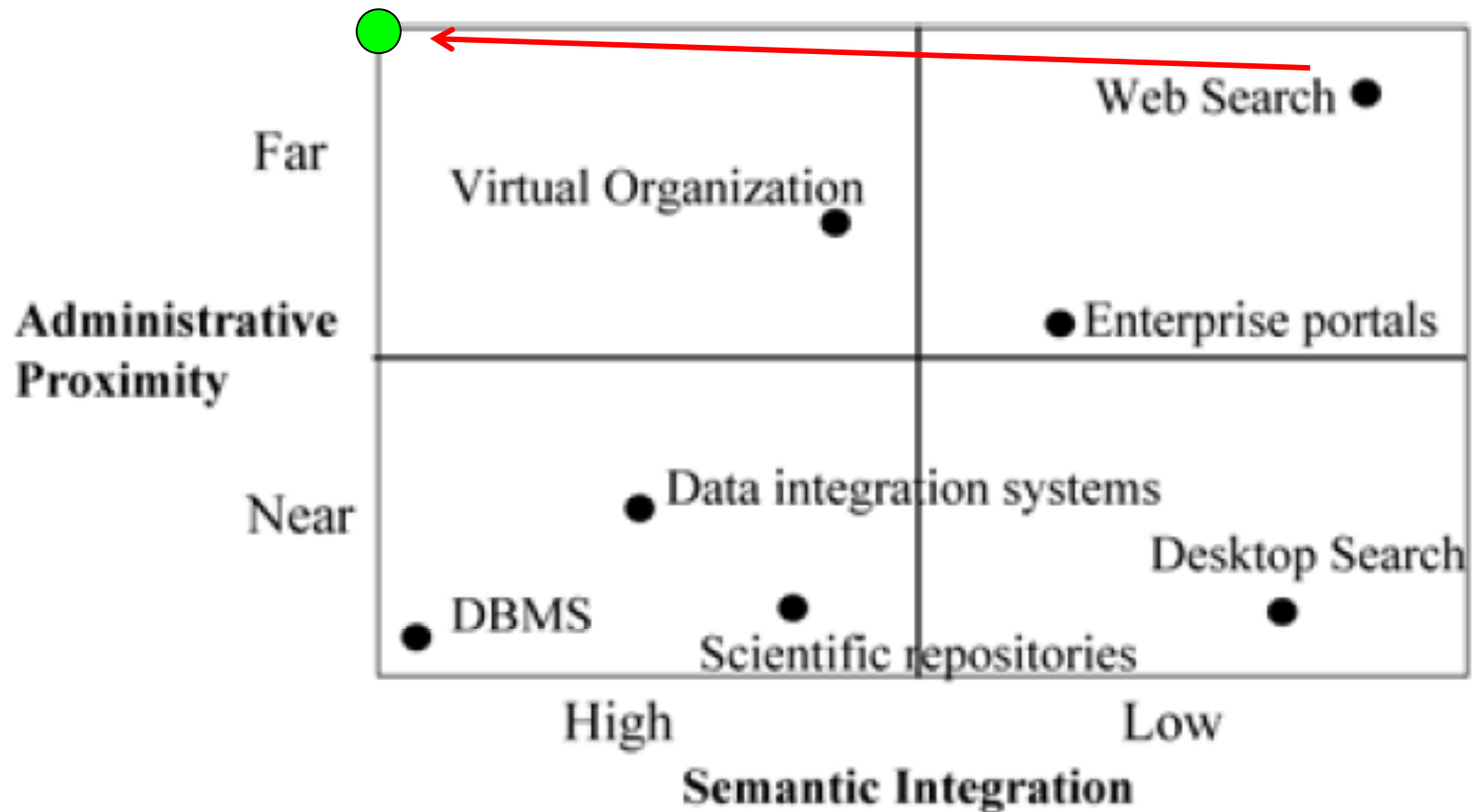
- a gradual process, making a bigger part of web content semantic
- increasing degree of automated web content processing
- third party semantization
- Yes, we know there is Google, Semantic web initiative, Web2.0, social web, ...

# Web Semantization – using Automated Web Information Extraction Tools

---



# Web Semantization – using Automated Web Information Extraction Tools



<http://portal.acm.org/citation.cfm?id=1107499.1107502>

# WIE – our approach – phases of annotation

## 1. Classification

Tabular page ~ Textual Page, Domain, etc.

## 2. General domain independent intermediate annotation

Different for tabular pages and textual pages

## 3. Domain dependent annotation

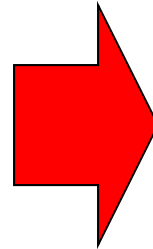


# Our Methods of Semantic Annotation

---

1

- Generally applicable
- Domain independent
- Intermediate
- Automated
- Created by **human experts**



2

- Domain dependent
- Adjusted by **unskilled user**

Two different approaches:  
**Textual pages**  
**Tabular pages**

# Our division of annotation tasks

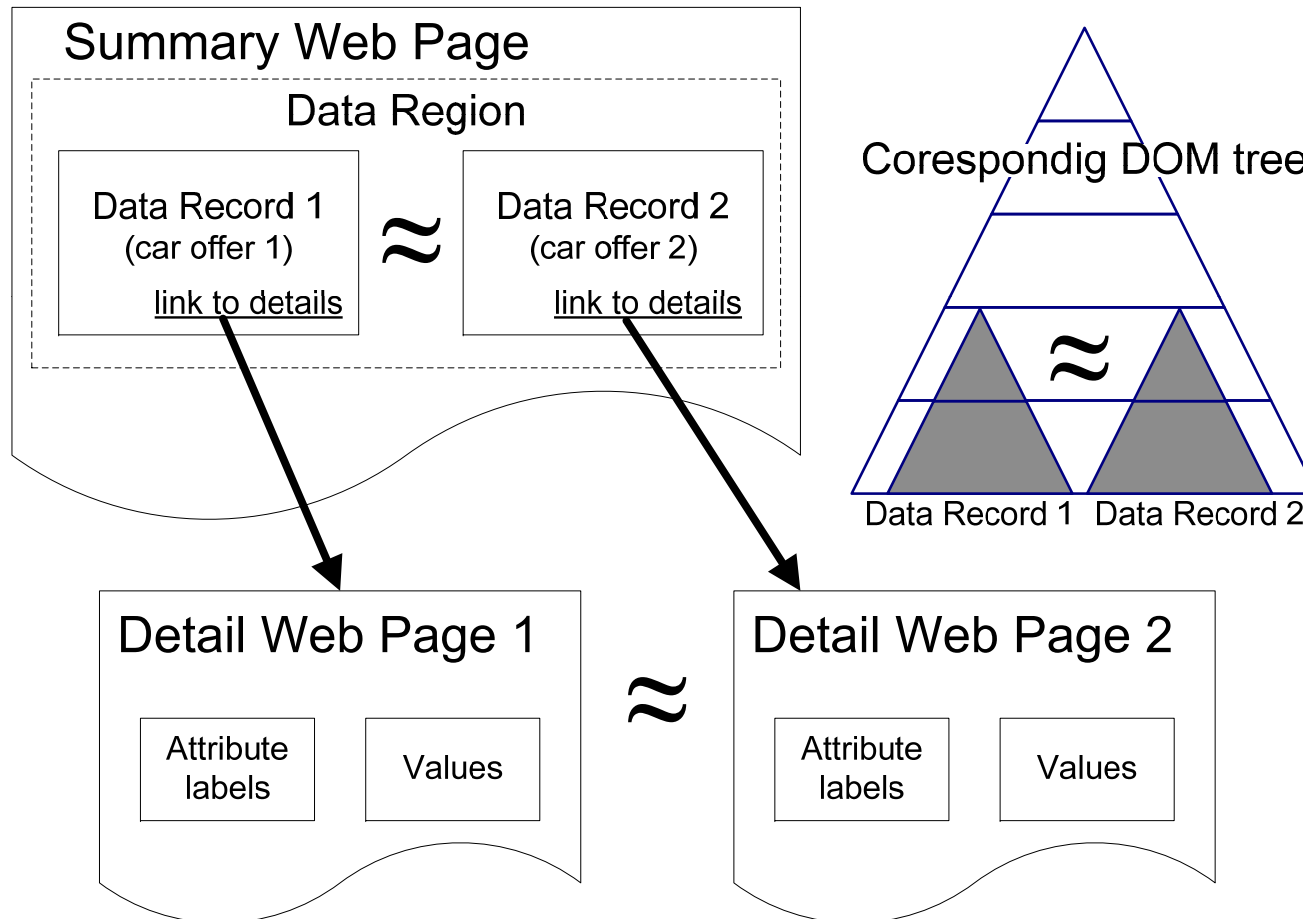
---

Type of annotation	Tabular pages	Textual pages
Intermediate DI - domain independent	Uses similarities	Does not use similarities
DD - Domain dependent	Does not use similarities	Uses similarities



# DI – intermediate annotation of tabular pages

---



# DI – ontology mining? work of R. Novotný

```
<div class="spacer"></div><br />
<div class="w98p">
  <div class="tblHeadGSm" id="headCat">
    <h1 class="font2R">
      FUJITSU-SIEMENS Amilo PRO V3205/ Core Duo T2350/
    </h1>
  </div>
</div>
<div class="w-cent">
  <form action="/Order/Order1.asp" method="post">
    <input type="hidden" name="NameItem" value="FUJITSU
  <table class="tblDetData">
    <tr>
      <td class="tblDDL">
        Vaše cena bez DPH:
      </td>
      <td class="fbold fblack fbig">
        25 129,00 Kč
      </td>
    </tr>
  </table>
  FUJITSU-SIEMENS Amilo PRO V3205/ Core Duo T2350/ 12.1
  DVD±RW +DL/ WiFi/ BT2.0/ XP Pr
```

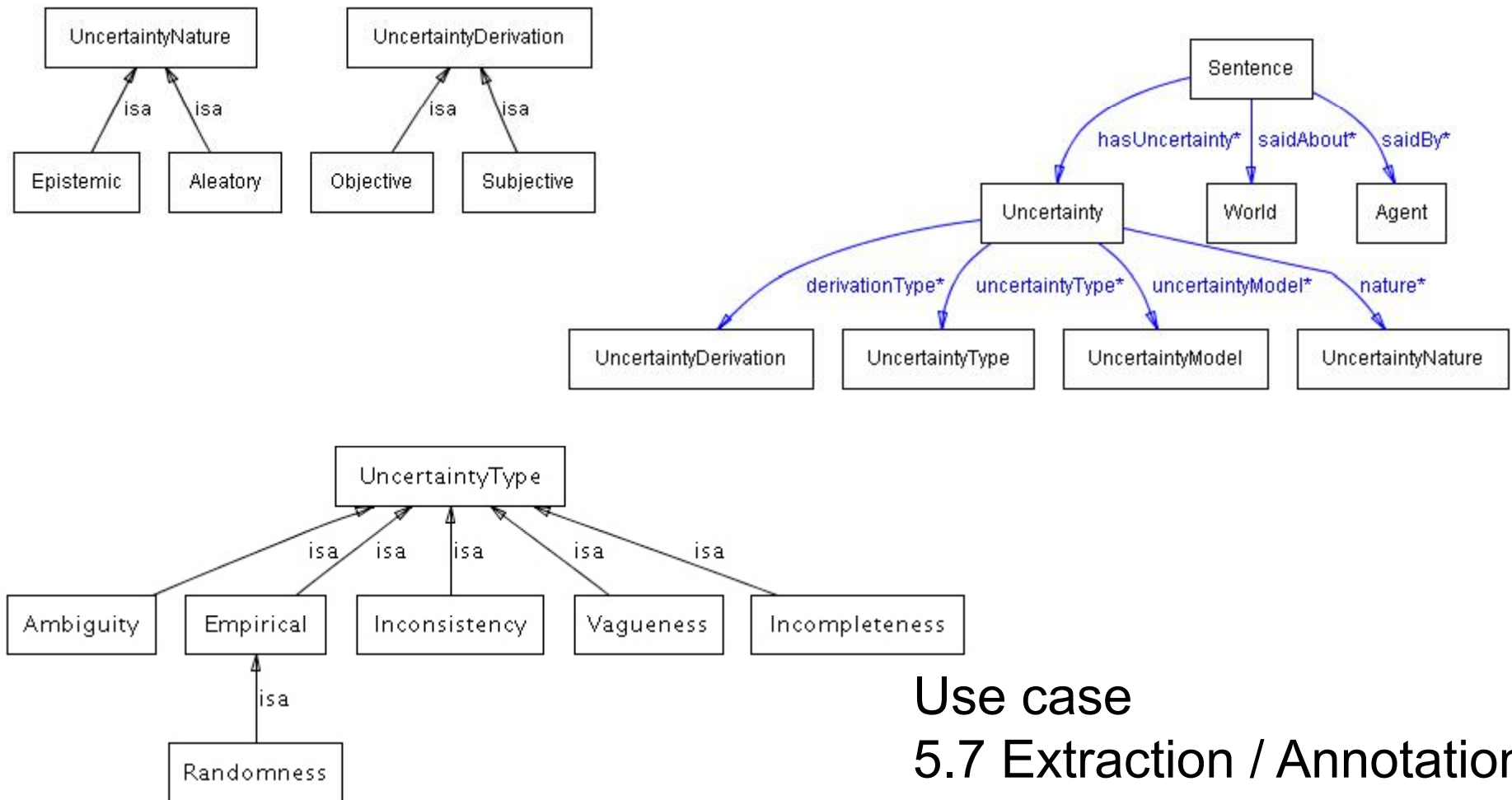
Vaše cena bez DPH:	25 129,00 Kč
Poplatek - autorská odměna:	50,00 Kč
Vaše cena bez DPH včetně poplatků:	25 179,00 Kč
Vaše cena včetně DPH:	29 963,01 Kč
Běžná cena - doporučená výrobcem:	<del>33 034,40 Kč</del>
Ušetříte - sleva z běžné ceny	9,30% / 3 071,39

```
<div class="spacer"></div><br />
<div class="w98p">
  <div class="tblHeadGSm" id="headCat">
    <h1 class="font2R">
      FUJITSU-SIEMENS Amilo PRO V3515/ C-M 430/ 15.4"
    </h1>
  </div>
</div>
<div class="w-cent">
  <form action="/Order/Order1.asp" method="post">
    <input type="hidden" name="NameItem" value="FUJIT
  <table class="tblDetData">
    <tr>
      <td class="tblDDL">
        Vaše cena bez DPH:
      </td>
      <td class="fbold fblack fbig">
        12 049,00 Kč
      </td>
    </tr>
  </table>
  FUJITSU-SIEMENS Amilo PRO V3515/ C-M 430/ 15.4" WXGA
  +DL/ WiFi/ Bez OS
```

Vaše cena bez DPH:	12 049,00 Kč
Poplatek - autorská odměna:	50,00 Kč
Vaše cena bez DPH včetně poplatků:	12 099,00 Kč
Vaše cena včetně DPH:	14 397,81 Kč
Běžná cena - doporučená výrobcem:	<del>15 807,96 Kč</del>

Vaše cena bez DPH včetně poplatků:

# Uncertainty Reasoning for the World Wide Web



Use case

5.7 Extraction / Annotation  
W3C Incubator Group

# Machine processing of texts – J. Dědek

Ministerstvo vnitra  
home navigace vyhledávání změna vzhledu

## Zpravodajství

Informace z resortu o tom, co se stalo, co se děje i co se připravuje

### HZS Jihomoravského kraje

Zubatého 1, 614 00 Brno, telefon 960 630 111,  
<http://www.firebrno.cz>  
Zpravodajství v roce 2008

15.05.2007

#### V trabantu zemřeli dva lidé

*K tragické nehodě dnes odpoledne hasiči vyjžděli na silnici z obce Česká do Kuřimi na Brněnsku.*

Nehoda byla operačnímu středisku HZS ohlášena ve 13.13 hodin a na místě zasahovala jednotka profesionálních hasičů ze stanice v Tišnově. Jednalo se o čelní srážku autobusu Karosa s vozidlem Trabant 501. Podle dostupných informací trabant jedoucí ve z Brna do Kuřimi zřejmě vyjel do protisměru, kde narazil do linkového autobusu dopravní

Hasiči udělali na vozidle protipožární opatření a po vyšetření a zadokumentování nehody dopravní policií vrak trabantu zaklesnutý pod autobusem pomocí lana odtrhli. Po odstranění střechy trabantu pak z kabiny vyprostili těla obou mužů. Obě vozidla – trabant i autobus, pak postupně odstranili na kraj vozovky a uvolnili tak jeden jízdní pruh. Únik provozních kapalin nebyl zjištěn. Po 16. hodině pomohli vrak trabantu naložit k odtahu a asistovali při odtažení autobusu. Po úklidu vozovky krátce před 16.30 hod. místo nehody předali policistům a ukončili zásah.

**odkazy**

Hasiči

- Generální ředitelství hl. m. Praha
- Jihočeský kraj
- Jihomoravský kraj
- Karlovarský kraj
- Královéhradecký kraj
- Liberecký kraj
- Moravskoslezský kraj
- Olomoucký kraj
- Pardubický kraj
- Plzeňský kraj
- Středočeský kraj
- Ústecký kraj
- kraj Vysočina
- Zlínský kraj

V této rubrice Zpravodajství

- Aktualizace stránek
- Archiv zpravodajství
- Bleskové zpravodajství RSS
- Boj proti korupci
- Digitální televize
- Hasiči
- Hlavní zprávy
- Ministerstvo
- Od dopisovatelů (neoficiální)
- Policie
- Regiony
- Servis nejen pro novináře
- Schengenská spolupráce
- WebEditorial

Na našem serveru v jiných rubrikách

- Aktuality Národního archivu

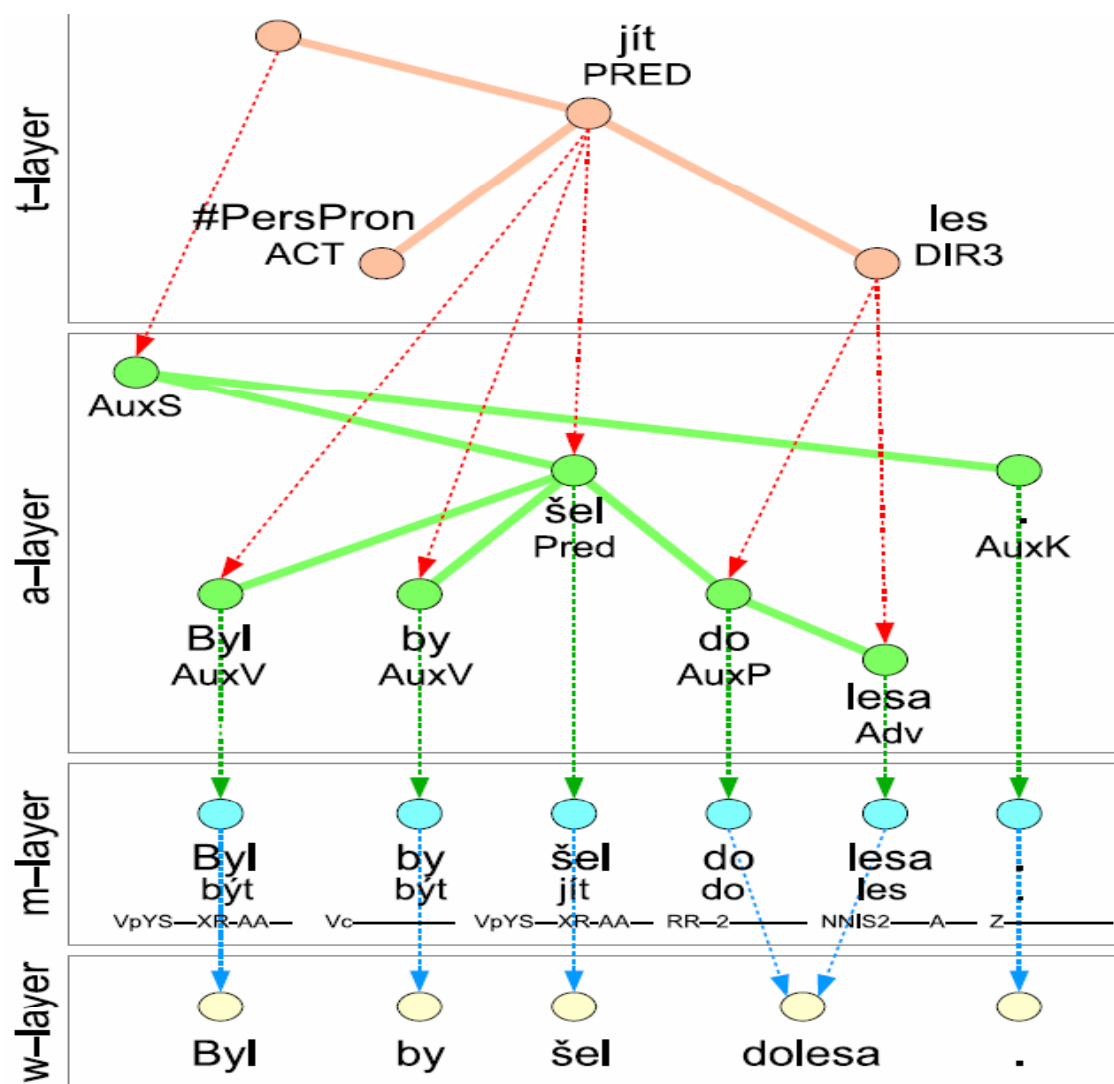
Dangerous roads  
detection,

How many people died,...

Which cars are dangerous?

...

# Dom. indep. – Linguistic annotation “UFALware”



Ministerstvo vnitra  
 Zpravodajství  
 Informace z resortu o tom, co se stalo, co se děje i co se připravuje

home navigace vyhledávání změna vzhledu

**HZS Jihomoravského kraje**  
 Zubatého 1, 614 00 Bmč telefon 950 530 111,  
 Http://www.firebrno.cz  
 Zpravodajství v roce 2006

15.05.2007

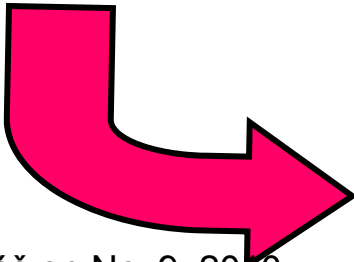
**V trabantu zemřeli dva lidé**  
 K tragické nehodě dnes odpoledne hasiči vyjžděli na silnici z obce Česká do Kuřimi na Brněnsku.

Nehoda byla dopravnímu středisku HZS ohlášena ve 13.13 hodin a na místě zasahovala jednotka profesionálních hasičů ze stanice Tisřov. Jednalo se o čelní srážku autobusu Karosa s vozidlem Trabant 601. Podle dostupných informací trabant jedoucí vo z Bmč do Kuřimi zřejmě vyjel do protisměru kde narazil do linkového autobusu dopravní společnosti ze Žďáru nad Sázavou. Ve zdemolovaném trabantu na místě zemřeli dva muži – 82letý senior a další muž, jehož totožnost zjišťují policisté.

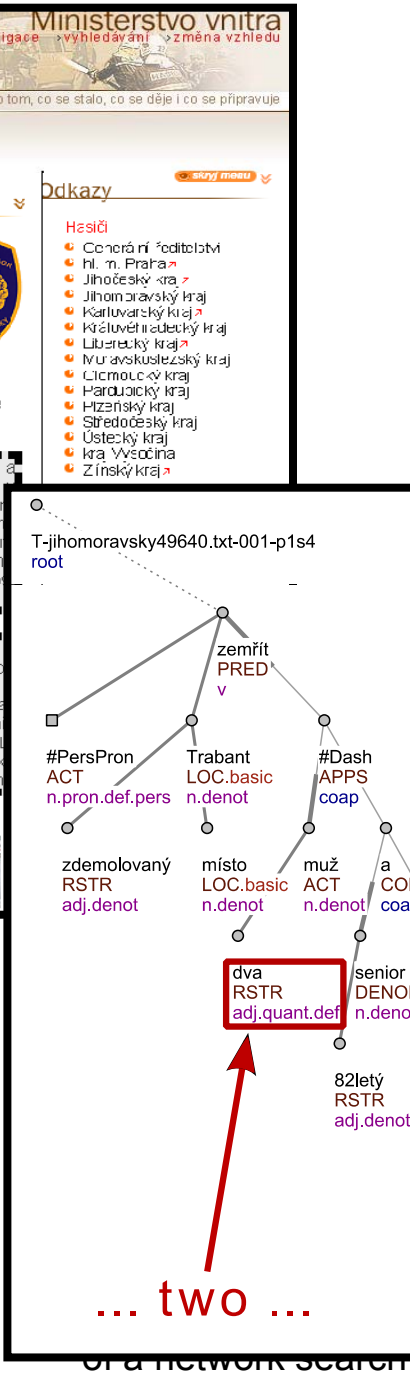
Hasiči udělali na vozidle protipožární opatření a po vyšetření zadokumentovali nehodu dopravní policii vrak trabantu zaklesnutý pod autobusem pomocí lara odtrhli. Po odstranění střechy trabantu oax kabiny vyprostili těla obou mužů. Čbě vozidla – trabant i autobus, pa postupně odstranili na kraj vyzvoxy a uvolnili tak jeden jzdni prun. Jn provozních kapalin nebyl zjištn. Po 16 hodině pomohl v vak trabant na ožít k odtažení a asistovali při odtažení autobusu. Po úklidu vozovky krátce před 16.30 hod místo nehody předal policistům a ukončil zásah



Source web page

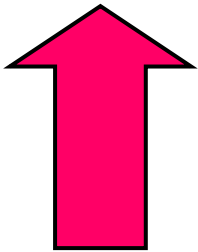


P. Vojtáš on Nov9, 2010 at LASSY, uni.lu



```
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



Linguistic trees

# Dom. dep. – ILP learning of linguistic rules

---

- Positive examples E+
  - examples with desired properties.
- Negative examples E-
  - examples with opposite properties.
- Background knowledge
  - additional data describing the examples and domain

## Rules found by ILP

```
injured(A) :- id(B,A),
id(B,t_plzensky57770_txt_001_p5
s2). Trivial
```

```
injured(A) :- id(B,A),
id(B,t_plzensky60375_txt_001_p1
s6). Trivial
```

```
injured(A) :- id(B,A),
edge(B,C), edge(C,D),
t_lemma(D,injure).
```

```
injured(A) :- id(B,A),
edge(B,C), edge(C,D),
t_lemma(D,accident).
```

# Conclusion of WIE and annotation part

---

Gradual semantization is possible!

## Goals

- minimalization of human assistance
- minimalization of human expert assistance

Work in progress, now just a „proof of concept“.

Key issue – scalability

**Software** available on request

Integration of UFALware with GATE

FILP, Vidome,...

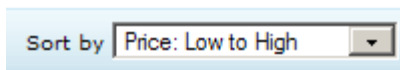
# User Preferences on the (Semantic) Web

---

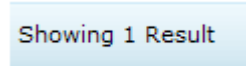
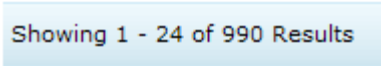
- Different users with different preferences
- So far no experiments with human users
- Focus on preference learning for visualization on web pages

# Motivation – current state

- Classical e-shop
  - Strict criteria on attributes
    - Price 50-100\$
  - Simple ordering by price, name



- Many or few objects







- Same answer for every user



# Motivation – dream state

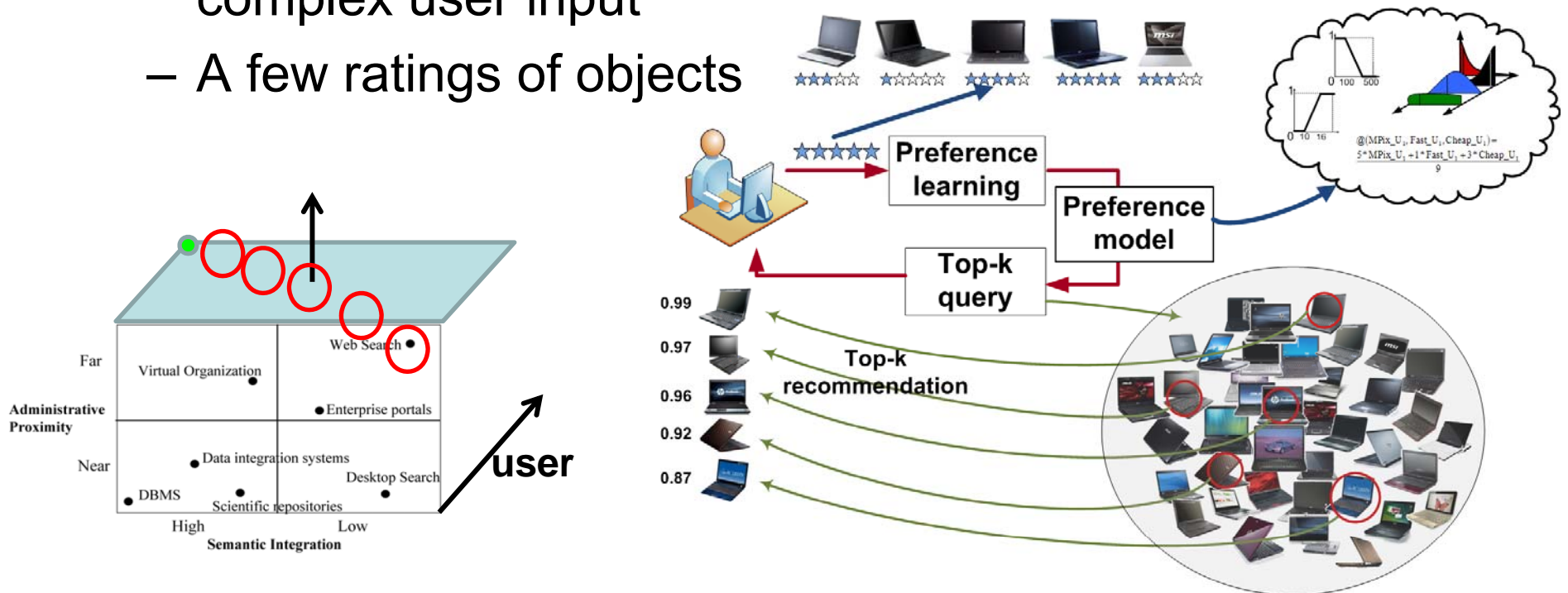
---

- Personalization
  - Results reflect the user, not only the query
- More insight for the user
  - Solve „no object“ or „too many objects“ problem
  - *How good is the notebook?*
    - Relaxed criteria
    - More complex ordering
- Ordering
  - Instead of restriction
- Specification of ideal attribute values
  - Rather than acceptable values

Notebook 251	96%	Notebook 188	59%
	If you're into high-def movies, games, music and photography, this laptop is for you. New Windows gives you more ways than ever of savoring your digital		Your on-the-go lifestyle might be hectic. But this laptop can smooth out the twists and turns. Powered by new Windows and reliable processing power from Intel
○ Manufacturer: ACER ● LCD size: 12 inches ● Wifi: Yes ● RAM size: 8192 MB > Processor: Core 2 Duo T7500		○ Manufacturer: ACER ● LCD size: 14.1 inches ● Wifi: Yes ● RAM size: 2048 MB > Processor: Core 2 Duo T7100	
● 38305,- incl. VAT 45582.95,-	In Stock 	● 28190,- incl. VAT 33546.1,-	In Stock 

# Motivation – preference learning

- Let's expect even less from the user
  - Instead of direct specification of ideal values - learning the most preferred values from a less complex user input
  - A few ratings of objects



# User model = Fagin's threshold algorithm model

---

- User model learning is divided into two steps

**1. Local preferences** - normalization of the attribute values of notebooks to their preference degrees

$$f_i : D_{A_i} \rightarrow [0,1]$$

Transforms the space  $\prod D_{A_i}$  into  $[0,1]^N$

- Manufacturer: ACER
- LCD size: 14.1 inches
- Wifi: Yes
- RAM size: 2048 MB
- > Processor: Core 2 Duo T7100

**2. Global preferences** - aggregation of preference degrees of attribute values into the predicted rating

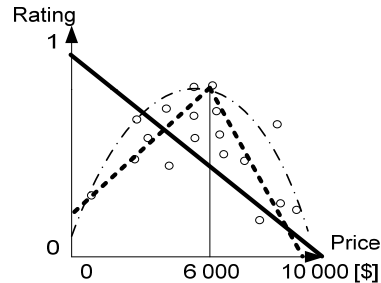
$$@ : [0,1]^N \rightarrow [0,1]$$

Notebook 188

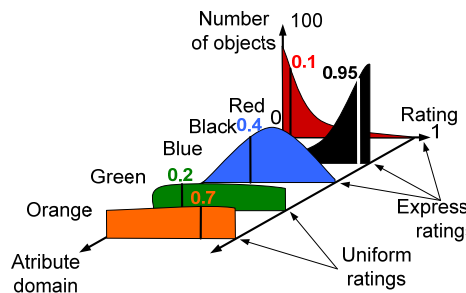
59%

# Learning user model – A. Eckhardt

- Local preferences
  - Regression for numerical attributes



- Average rating for nominal attributes



- Global preferences

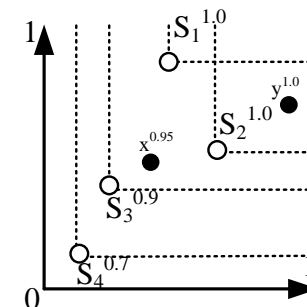
- *Statistical*

- Learned weights for weighted average

$$\begin{aligned} @(\text{RAM}_{U_1}, \text{CPU}_{U_1}, \text{Price}_{U_1}) = \\ \frac{5 * \text{RAM}_{U_1} + 1 * \text{CPU}_{U_1} + 3 * \text{Price}_{U_1}}{9} \end{aligned}$$

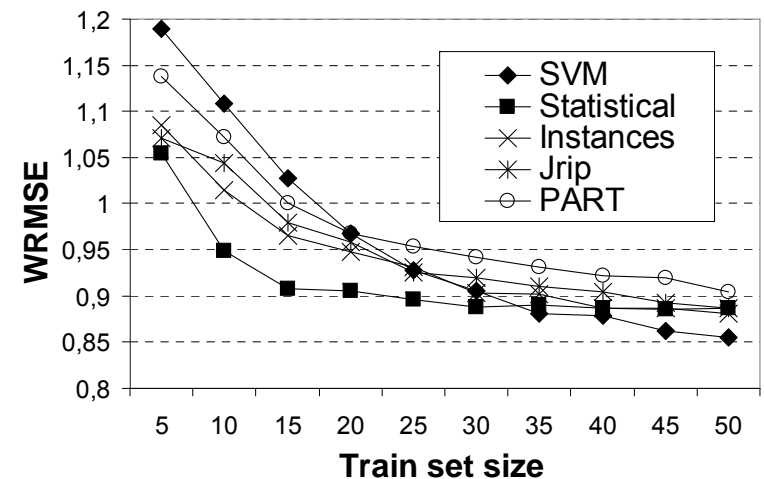
- *Instances*

- Uses objects from training set for estimation of rating



# Specific problems to preference learning

- Small training set
  - Users do not want to invest too much effort
- Transparent preference model needed - Why is the object recommended by the system
  - For using the information in user interface
- Different error measures than “standard” machine learning
  - Correct ordering is important, better rated objects are more important
  - WRMSE
    - $\sqrt{\sum_{o \in X} r(o)(\hat{r}(o) - r(o))^2 / \sum_{o \in X} r(o)}$
  - WTau coefficient
    - Compares two ordered lists and emphasizes better rated objects
  - Top-k score
    - Percentage of correct objects in top-k (without order)
  - Comparison with SVM, Perceptron,...



# Conclusion of user preference learning part

---

Preference learning can be useful.

## Goals

- minimalization of effort of the user
- maximalization of benefit for the user

Work in progress, now just a „proof of concept“.

Needed experiments of user interface with real users.

**Software** available at <http://code.google.com/p/prefwork/>  
for user interface, top-k querying and preference learning

# Conclusions

---

Web semantization – proof of concept

Integrated components

- WIE + annotation
- Repository
- User preference learning

Future?

# Department of Software Engineering

<b>People:</b>	14 full-time research employees 12 internal Ph.D. students 9 external employees
<b>Average results per year:</b>	80 papers in refereed proceedings/journals 7 SW prototypes
<b>Awards:</b>	8 best paper awards from international events
<b>Grants:</b>	10 co-operation team grants • Grant Agency of the Czech Republic 13 individual grants • Grant Agency of the Czech Republic • Grant Agency of the Charles University 5 development projects • Ministry of Education, Youth and Sports of the Czech Republic
<b>Cooperation:</b>	Organization of 16 international and 4 local conferences and workshops Cooperation with 12 EU, 7 non-EU and 5 local institutions Cooperation with 5 industrial partners
<b>Teaching:</b>	73 courses and subjects

# Research groups

---



## XML Research Group

<http://www.ksi.mff.cuni.cz/xrg/>

The XML Research Group (XRG) focuses on various aspects of XML technologies such as modelling of XML data and XML applications, statistical analysis of real-world XML data, inference of XML schemas, storage strategies for XML data, evolution and change management of XML data and XML applications, similarity of XML data, XML benchmarking etc.



## Siret Research Group

<http://siret.ms.mff.cuni.cz/>

The main interests of Siret research group are the similarity search in multimedia databases, protein similarity retrieval, similarity modelling, database indexing - metric access methods, shape extraction & image retrieval.



## Web Semantization Research Group

<http://www.ksi.mff.cuni.cz/semwex/>

The Web Semantization Research Group (SEMWeX) focuses on various aspects of semantic web technologies. The research is concentrated primarily on dissemination of the idea of global semantization, profounding the usage of XML technologies, interconnecting seemingly disjoint research areas and experimental evaluation of formal and theoretic results.

# Research groups

---

The logo for the Service-Oriented Systems Group (SOSG) features the letters 'SOSG' in a bold, yellow, sans-serif font. The letters are set against a dark, circular background that has a glowing, fiery effect around the edges.

## Service-Oriented Systems Group

<http://www.ksi.mff.cuni.cz/sosg/>

- agile SOA, dynamic service networks, and flexible information and control systems (flexible manufacturing, ...)
- data and information quality issues in health care, education quality evaluation, and state administrative in general
- adaptive language technologies (using compiler techniques in dynamic communication systems)
- application of formal methods in requirements specification, modeling, and system design



## Faculty of Mathematics and Physics

<http://www.mff.cuni.cz/toUTF8.en/>

## School of Computer Science

<http://cs.mff.cuni.cz/>

Thank you for your attention!

Questions?