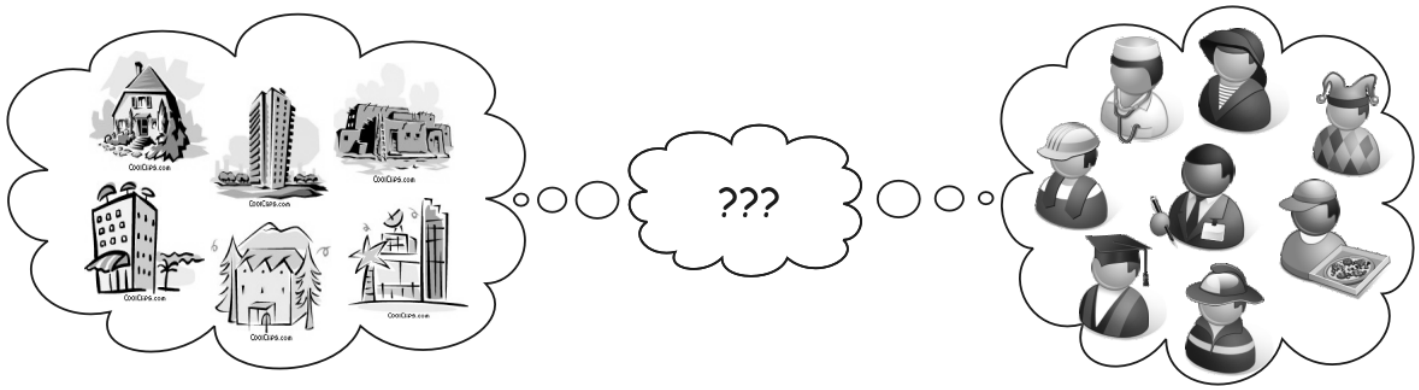
# Towards supporting interaction of user and web

Leo Galamboš, Peter Vojtáš,
Jakub Yaghob, Filip Zavoral

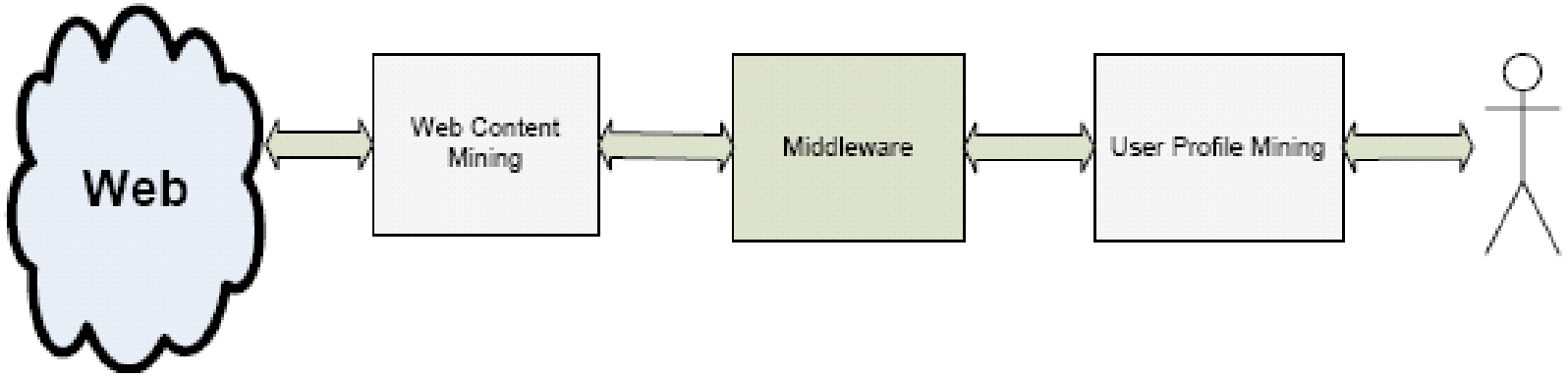
MFF UK Praha

# Motivace




- Mnoho objektů zájmu (např. hotely)
- Mnoho uživatelů s různými preferecemi

# Motivace




- Řešení jako řetěz nástrojů
- Ne proprietární řešení ale s formálními modely v pozadí
- Přenesitelné do různých domén
- Míry kvality (>1TB, 0.5 s, přesnost, úplnost, ...)

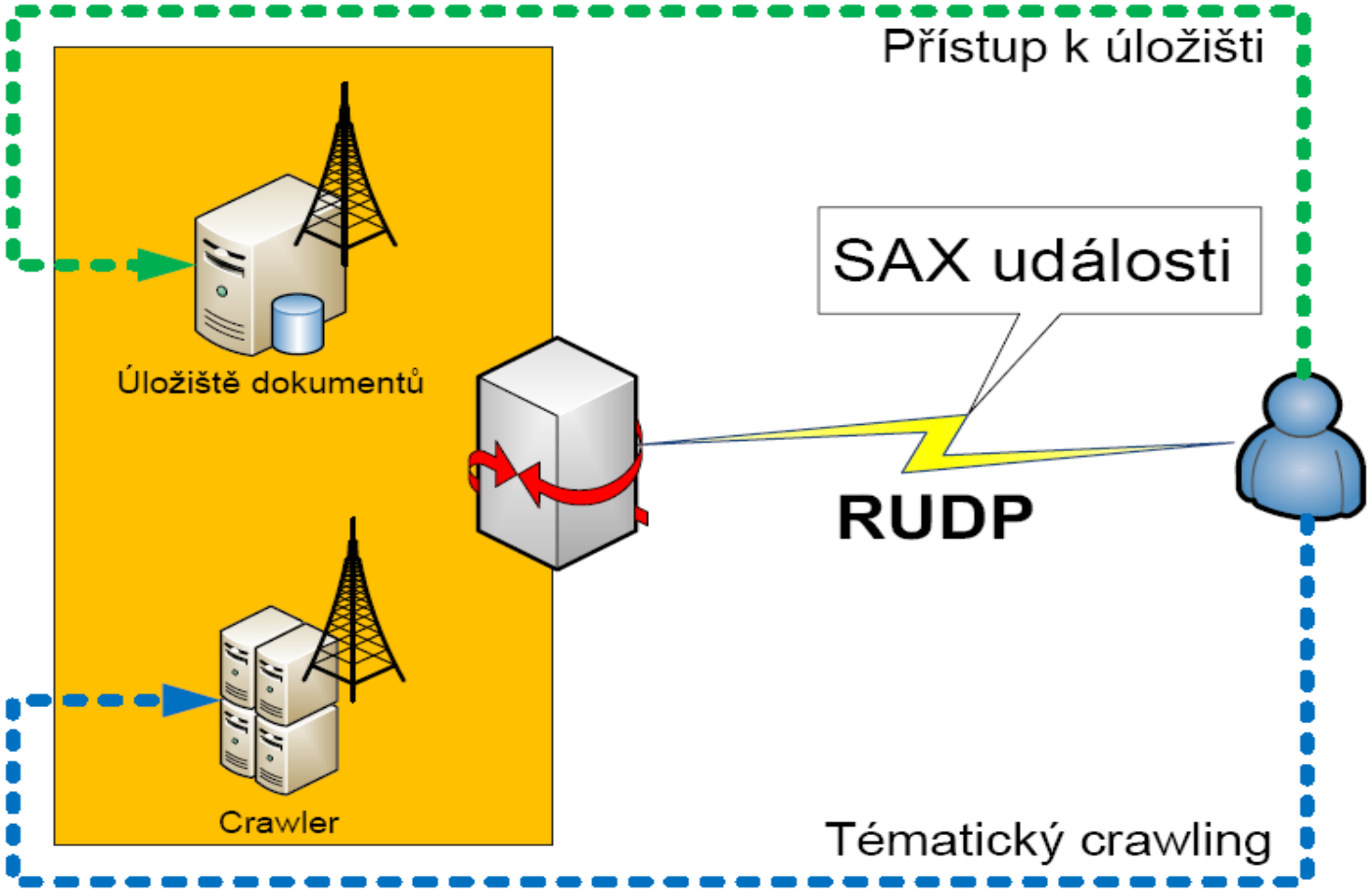
# Přehled

- Crawler Egothor (LG)
- Infrastruktura Trisolda (JY+FZ)
- Wrapper Vidome (DP D. Maruščák)
- Anotace textových zdrojů (DP J. Dědek)
- Modely uživatelských preferencí
- Top-k dotazů TOKAF (DP A. Eckhardt)
- Otevřené problémy
- Závěr

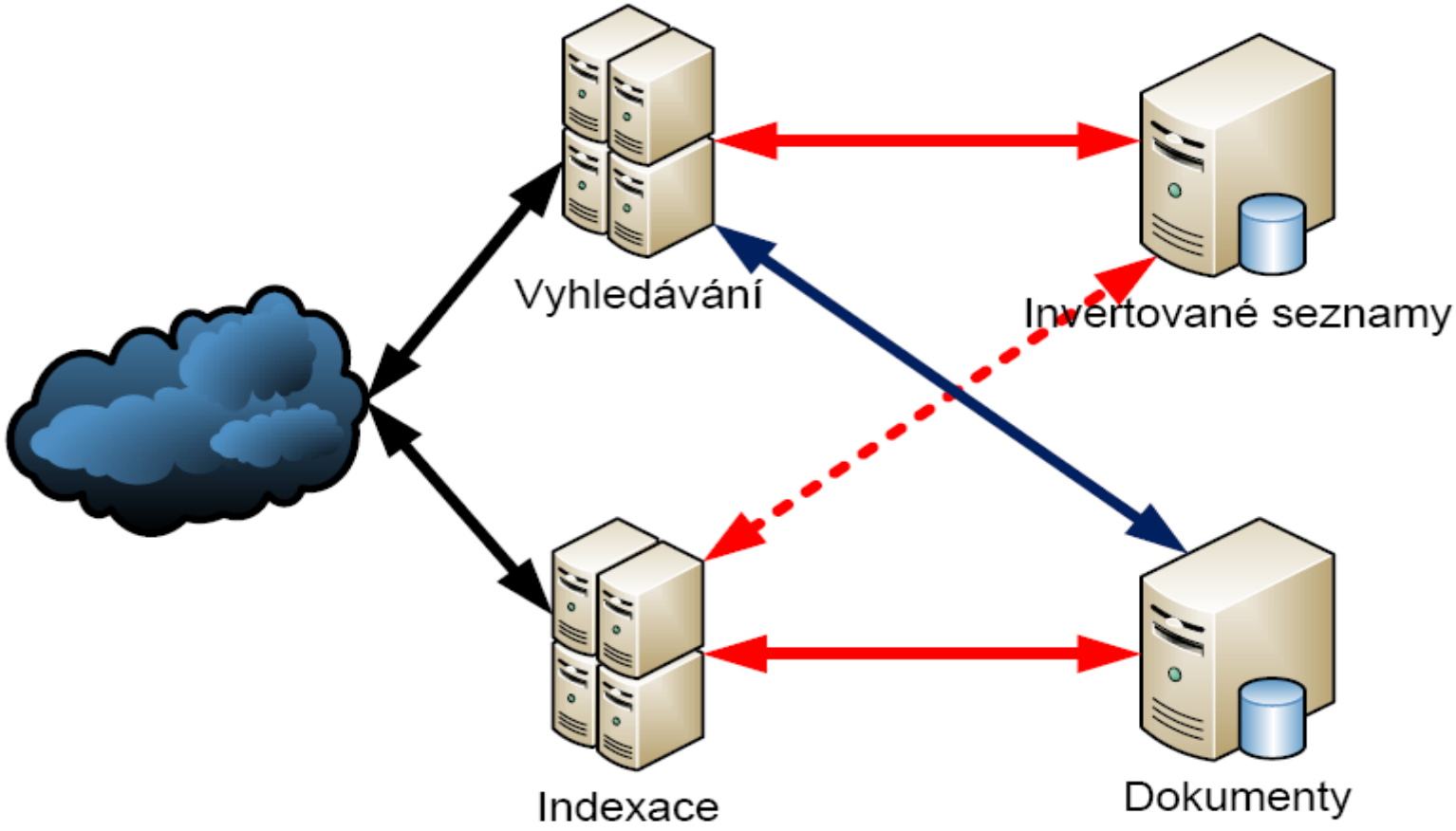
# Crawler Egothor

Leo Galambos

# Co chceme pro projekt sémantického webu



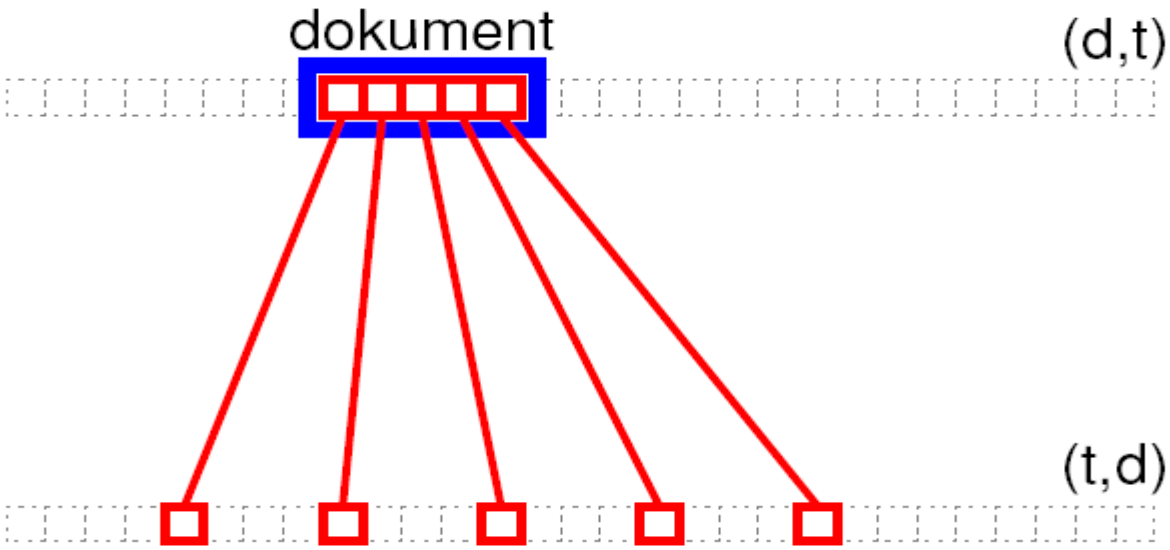
# Co chce praxe po systému



# Otázky

- rychlost crawlování: závisí kvalita indexu (vyhledávání) na technice crawlování?
- jaká zátěž jde na datové struktury při správném crawlování?
- jak spravovat kompaktní index?
- vyřeší něco keš nad dotazy a odpověďmi WVS?
- jak konstruovat úložiště?

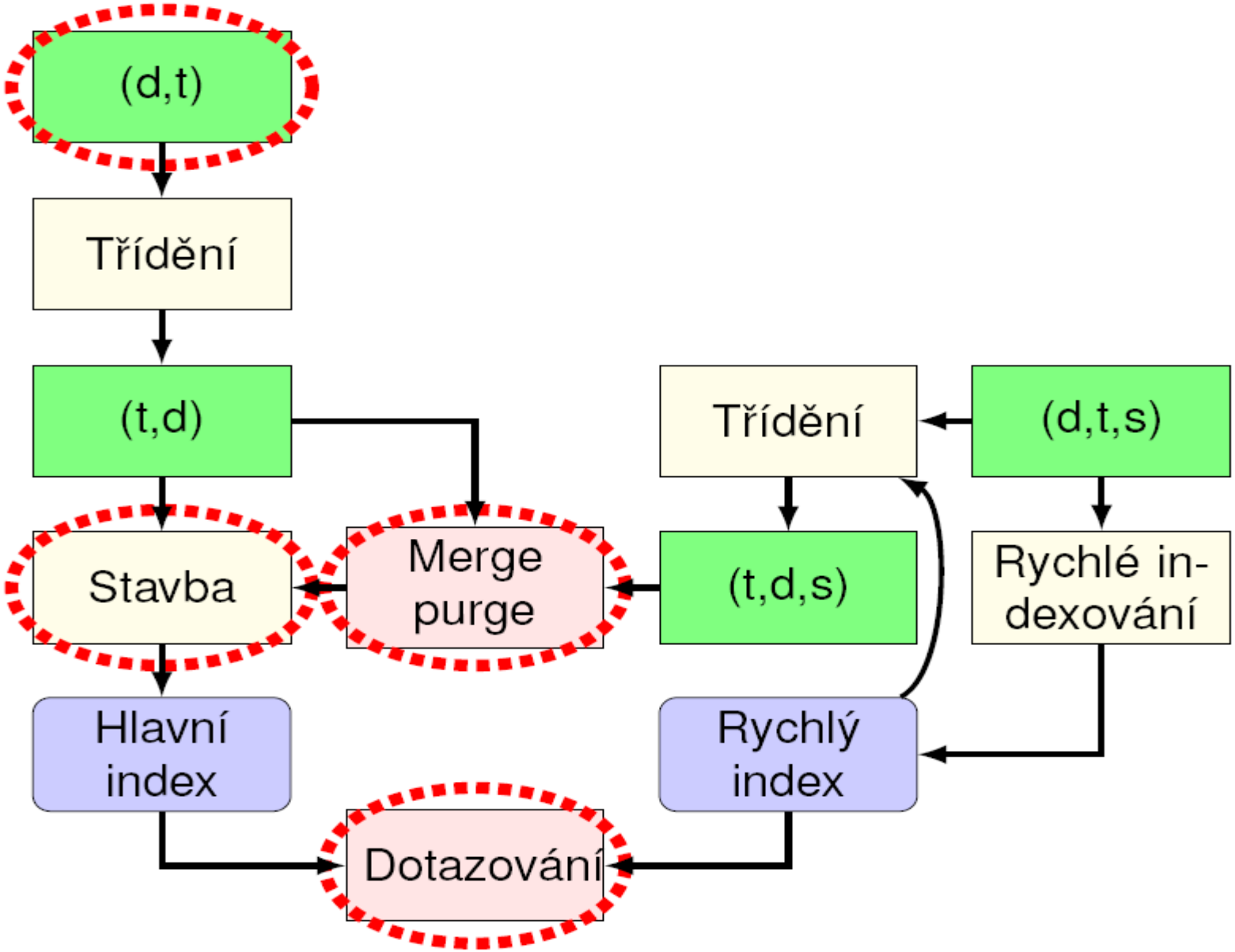
# Co se děje s daty




## Produkt

- zápor: aktualizace, výmaz, vložení dokumentu
- klad: rychlé řešení dotazů

# Webový vyhledávací systém



# Problematická místa

- DNS (Mercator) – via Indexer-Async DNS
- test na (ne)existenci URL ve WVS
- duplikáty (.pt doména)
  - -36%: index.htm(l)
  - -30%: egothor.org v. www.egothor.org
  - -5%: /dir v. /dir/
  - +26%: auto-redirect
- SPAM. . .

# Výkony robotů

| Robot | #strojů/CPU | str/sec | tok (MB/s) | rok |
|---|---|---|---|---|
| Googlebot | 4/? | 25-32 | 0,2 | 1998 |
| Mercator | 1/2 | 55 | 0,84 | 1999 |
| Mercator | 4/8 | 72 | 1,72 | 2001 |
| Xyro | 4/? | 12 | ? | 2001 |
| Nutch | 1/? | ? | 0,5 | 2004 |
| Become | 50/? | 100 | 0,3? | 2004 |
| Egothor 1 | 1/2 | 40 | 0,6 | 2004 |
| Dominos | 5/? | 154 | 0,9 | 2004 |
| Egothor 2 | 1/2 | 70 | 1,5 | 2005 |
| Larbin | 1/2 | 80 | 2,1 | 2006 |
| Egothor 2 | 1/2 | 100 | 5,0 | 2007 |

# Existující řešení aktualizace indexu

1. znovu od základu – velké I/O ztráty
2. B-tree (Cutting, Pedersen) – velká fragmentace
3. $\Delta$-změny, pouze append – snadná implementace, velké nároky na filesystém i O/S
4. Forward Index (Brin, Page) – velké I/O ztráty
5. Landmarks (IBM, Google) – náročné na CPU i seek-y
6. Dynamizace (LG)

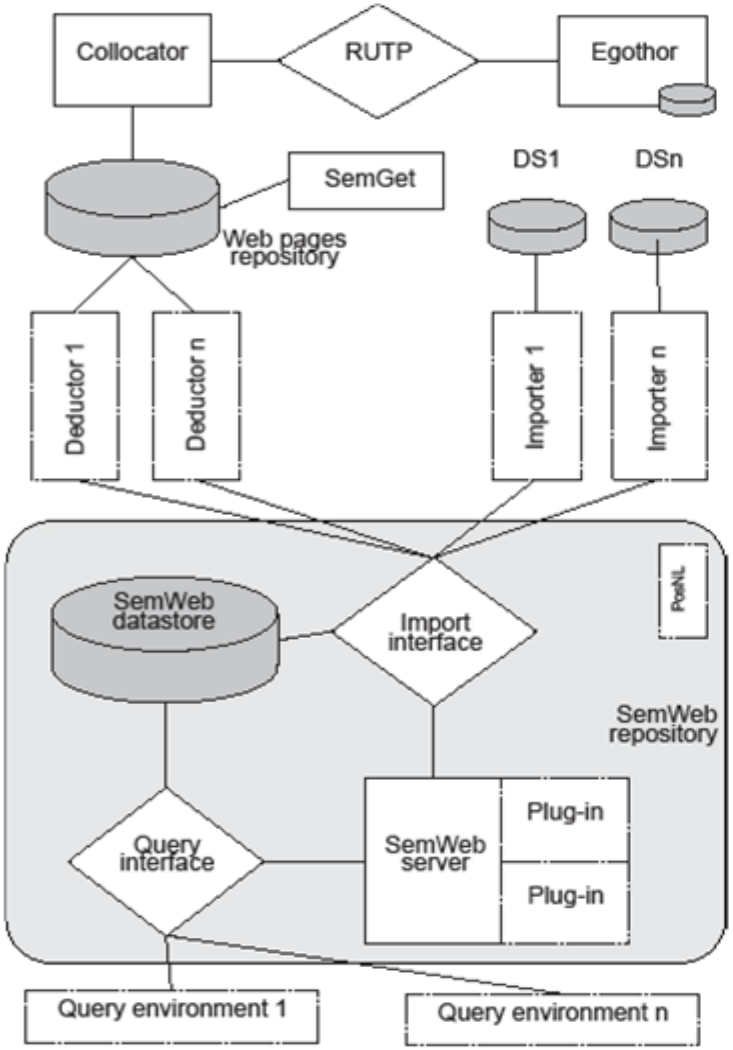
# Závěr

- plánovač robota může ovlivnit kvalitu výsledků - ano, výrazně
- hrubá síla versus komplikované plánování - hrubá síla, ale opatrně s výběrem plánovacího algoritmu
- dynamizace zajišťuje efektivnější správu než append-only
- keš v *brokeru* výrazně nepomůže
- potřebujeme úložiště s výkonem kolem 10000 doc/sec (neexistuje, potřeba vlastního vývoje)

# Infrastruktura Trisolda

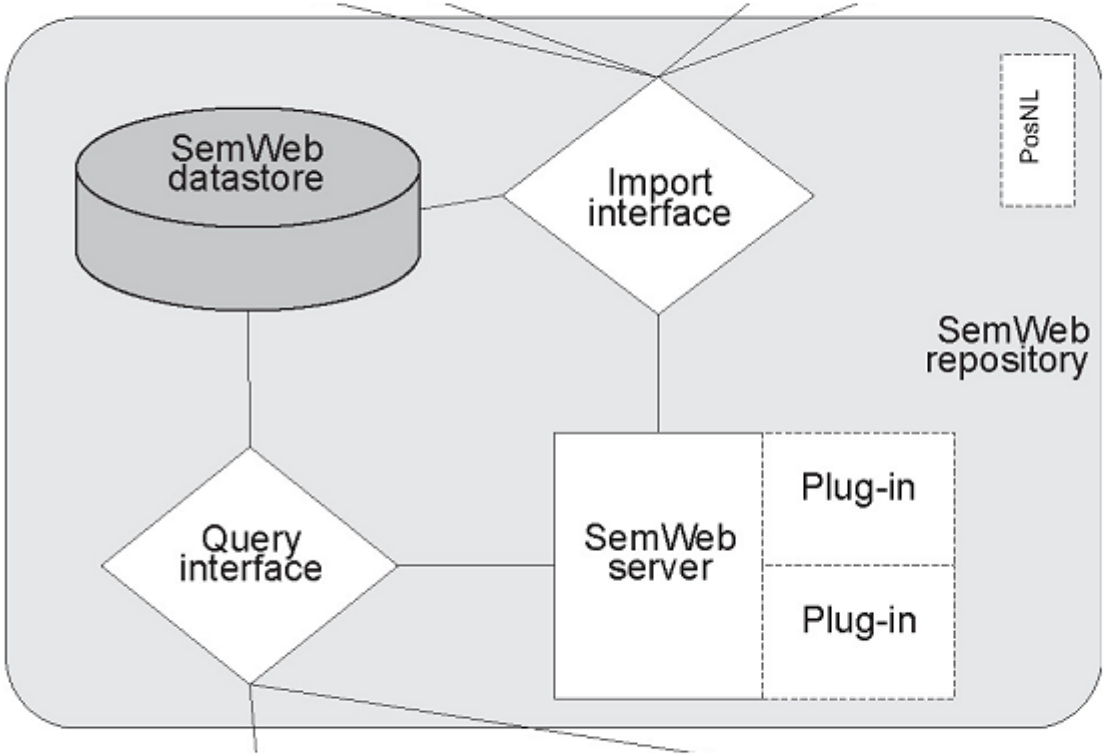
Jakub Yaghob a Filip Zavoral

# Trisolda overview




- Import paths
- SemWeb repository
- Query environments

# Trisolda repository



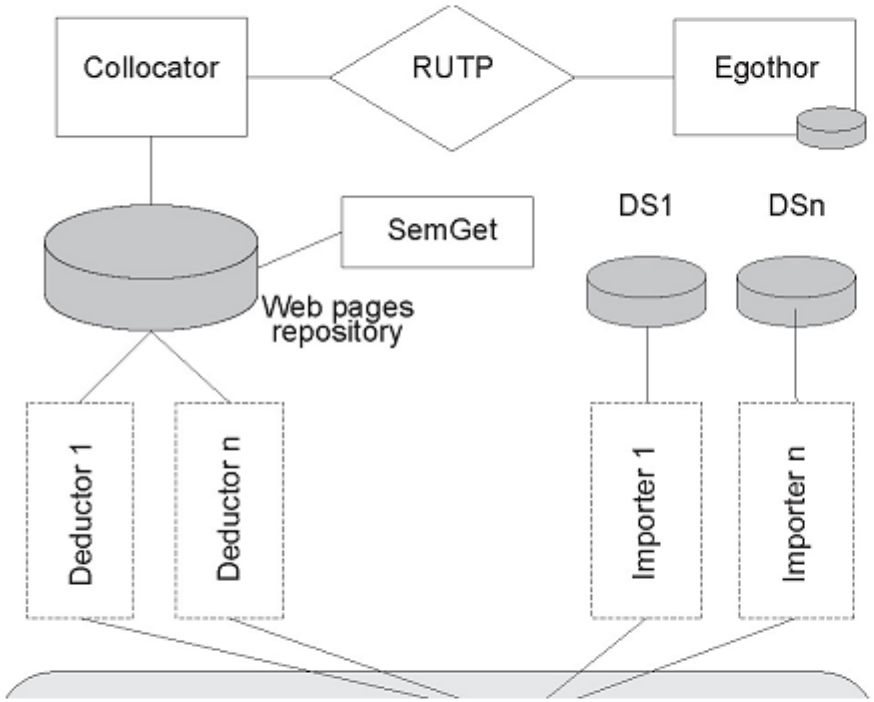

- Stores incoming data
- Retrieves results for queries
- Stores used ontology

- Import interface, query interface, application server
- Trisolda datastore – holds data in any format

# Trisolda application server

- Not all accurate data and inferred knowledge available at the moment of data import
  - The knowledge of the world is not accurate
- Background worker
  - Inferencing, data unifications, reasoner
  - Uses import and query APIs
  - Framework for server's plug-ins
    - Other experimental implementations of reasoners, unifiers, etc.

# Import paths



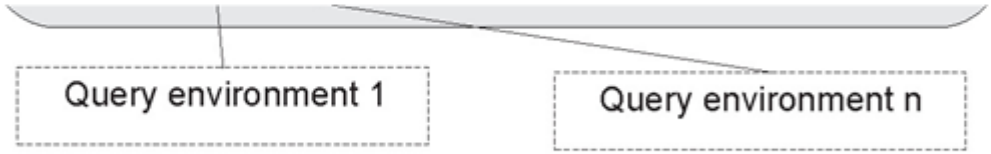

- Batch insert
- Immediate insert

- Direct import
  - data in data sources
  - converters to the used ontology
- Crawling wild Web
  - Egothor web crawler
  - parsed pages stored
  - deductors deduce data and its ontology

# Query API

- Based on simple graph matching and relational algebra
- Simple graph matching – SPARQL-like
  - Query: set of RDF triples with variables
  - Result: multiset of possible variable mapping – a relation (table)
- (the good old) relational algebra
- Not another SQL-like language
  - set of C++ classes and operators
  - query API used by software
- Query evaluation
  - different level of support by storage engines

# Query environments




- Q.E. present outputs from the repository
  - Using Query API
  - Examples: SPARQL compiler, repository browser, RDF visualizer, semantic executors, …

# Conclusions & future work

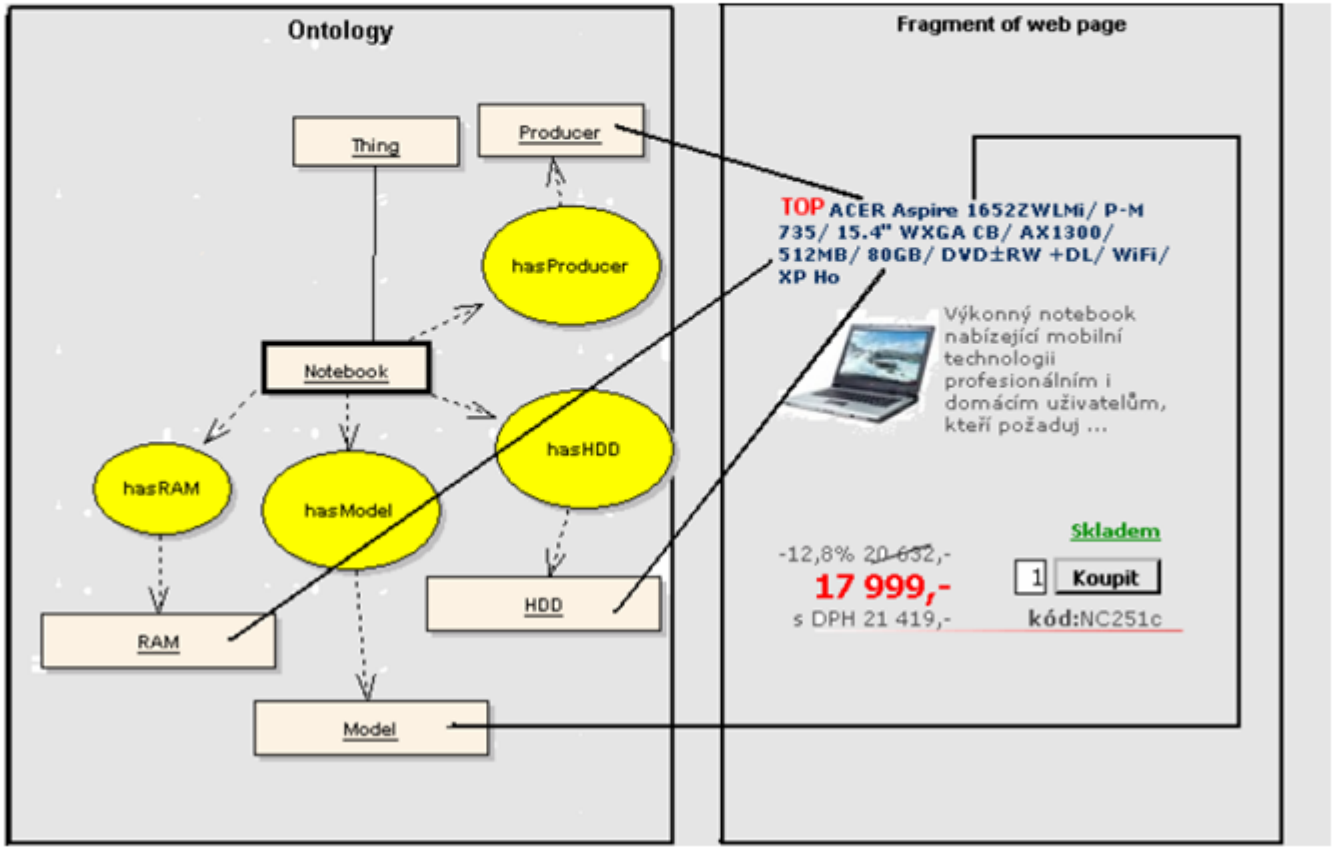
- Working infrastructure
  - gathering, storing and querying of semantic data
  - platform for future semantic web research
- Long-term goals
  - specialized semantic data storage
  - interface-based loosely coupled network of Semantic Web repositories
  - semantic computing, semantic services

# Wrapper Vidome

Diplomová práce

Dušan Maruščák

# Ontology based attribute value extraction

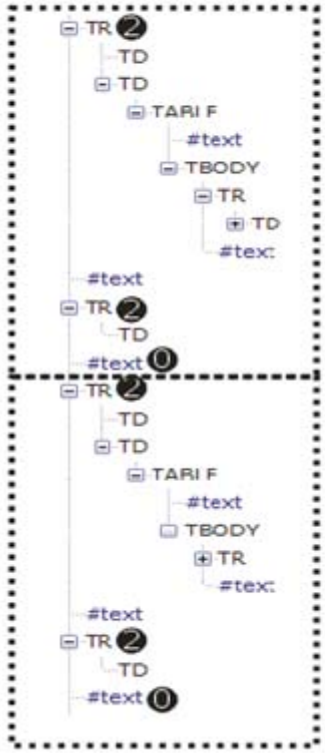


# Problémy


Nejprodávanější zboží této kategorie:
1. HP PAVILION dv6019ea/ Core Duo T2080/ N7400/ 15.4" WXGA BV/ 1GB/ SATA 120GB 5.4k/ DVD±RW +DL-RAM/ WiFi/ CAM/ VIS HP 25 597,00
2. HP COMPAQ nx7400/ C-M 430/ 15.4" WXGA BV/ 512MB/ 120GB 5.4k/DVD±RW +DL-RAM/ WiFi/ Bez OS 14 568,00
3. FUJITSU-SIEMENS Amilo Pa1538/ Turion 64 X2 TL52/ 120GB 5.4k / N7400/ 1GB/ SATA 160GB/ DVD±RW +DL/ WiFi/ VIS HP 22 248,00
4. ACER Aspire 3693WLMi/ C-M 430/ 15.4" WXGA CB/ 1GB/ 120GB 5.4k/ DVD±RW +DL-RAM/ WiFi/ CAM/ XP Ho 19 042,00
5. ASUS F3TC-AP060/ Turion64 MK36/ 15.4" WXGA CS/ N7300/ 1GB/ SATA 120GB 5.4k/ DVD±RW +DL-RAM/ WiFi/ BT/ CAM/ Bez OS 21 279,00
TOP FUJITSU-SIEMENS Amilo PRO V3205/ Core Duo T2350/ 12.1" WXGA CV/ 1GB/ SATA 120GB 5.4k/ DVD±RW +DL/ WiFi/ BT2.0/ XP
A
Miniaturní elegantní notebook v jehož útrobách nalezname výkonný dvoujádrový procesor firmy Intel ...
B
C
Skladem > 5 ks
-9,3% 33 206,-
30 118,-
s DPH 35 841,-
1 Koupit
kód:ND032e
TOP ASUS F3TC-AP008/ Turion64 X2 TL52/ 15.4" WXGA CS/ N7300/ 1GB/ SATA 120GB 5.4k/ DVD±RW +DL-RAM/ WiFi/ BT/ CAM/ Bez OS
Moderní výkonný notebook osazený dvoujádrovým procesorem AMD Turion64 X2 TL52 s frekvencí 1,6 GHz ...
Skladem 1-2 ks
-13,8% 27 201,-
23 444,-
s DPH 27 898,-
1 Koupit
kód:NA424g
TOP MSI MegaBook L745-023CZ/ Core 2 Duo T5600/ 17" WXGA/ N7600/ 1GB/ 120GB 5.4k/ DVD±RW+DL/ WiFi/ BT/ CAM/ Bez OS + ZDARMA Promo MSI StarReader - externí USB2.0 čtečka 52v1 pro paměťové karty CF/ MD/ MS/ MSPRO/ MSDUO/ MMC/ RS-MMC/ SD/ SIM
Spolehlivý výkonný notebook v elegantním provedení přináší perfektní zážitek při sledování videa ...
Skladem 1-2 ks
-21,4% 42 806,-
33 013,-
s DPH 39 286,-
1 Koupit
kód:NB026
TOP ACER Aspire 1652WLMi/ P-M 735/ 15.4" WXGA CB/ AX1300/ 512MB/ 80GB/ DVD±RW +DL/ WiFi/ XP Ho
Výkonný notebook nabízející mobilní technologii profesionálním i domácím uživatelům, kteří požaduj ...
TOP IBM/LENOVO THINKPAD T43/ P-M 760/ 14.1" SXGA+/ AX300/ 512MB/ 80GB 7.2k/ DVD±RW +DL/ WiFi/ BT/ FPR/ XP Pr
Velmi odolný a spolehlivý notebook určený pro náročné uživatele požadující vysoký výkon při nízké ...
TOP NEW ASUS F3JR-AP120C/ Core Duo T2080/ 15.4" WXGA CS/ AX2300/ 2GB/ SATA 120GB 5.4k/ DVD±RW +DL-RAM/ WiFi/ BT2.0/ CAM/ VIS HP + ZDARMA KWORLD 323U - analogový TV / digitální DVB-T tuner, externí USB2.0 dongle, software, dálk.ovládání, anténa
Moderní spolehlivý notebook, který Vás jistě zaujme svou velmi příznivou cenou.
• ACER Aspire 3694 C-M 440/ 15.4" WX CB/ 512MB/...
• ASUS F3JR-AP075 Duo T2350/ 15.4" W CS/ AX2300/...


Search results
Type your dates here, then click 'Go' to check availability
Booking online is a snap – or call 1(800) 608-2013 to book by phone
Hotel list view
Area map view
Hotel map view
Show hotels in this area:
London (and vicinity: (All areas)
Go
Hotel amenities:
Refine your search
Sort by:
Expedia Picks
Hotel Name
City
Hotel Class
Traveler Ratings
The Grand at Trafalgar Square
Restored Victorian hotel built in 1867 and situated just off Trafalgar Square in the centre of London; the National Gallery, Covent Garden and theatre ...
Traveler Opinion
4.4 out of 5
30+ reviews
Book online or call 1-800-451-2409 to speak with an Expedia agent
Enjoy rates from $176.67 per night between Jun-2-2007 and Aug-1-2007.
Availability & prices
The Strand Palace
Imposing Art Deco hotel in the theater district
Inhabiting its own block on the busy Strand, this hotel in ten blocks from Covent Garden, three blocks from the River Thames, and a half-km (a quarter mile) ...
3.4 out of 5
Data Region
Availability & prices
The Cumberland - a Guoman Hotel
Contemporary decor overlooking Hyde Park
Directly opposite Marble Arch, the eight-story Cumberland is across the street from Hyde Park, a block from Oxford Street shopping, and next door to an ...
4.4 out of 5
Book online or call 1-800-451-2409 to speak with an Expedia agent
Enjoy rates from $236.22 per night between Jun-2-2007 and Aug-1-2007.
Availability & prices
Club Quarters St. Pauls
Located in the courtyard of historic Stationers Hall, this restricted-entry hotel is adjacent to St. Paul's Cathedral.
4.1 out of 5
Data Record
Data Record
Data Record

# Řešení




Search over DOM form of page

Non-contiguous records


Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/Notebook/EXTRACT/http://www.alza.cz/18842920/notebooky.htm
VidomeXPI
Extractor Ontology Server
Mine instance from >>
Show instance
Clear instance
About
VIDOMEXPI
User interface for ontology mir
Notebooky | alza.cz
http://localhos...0/notebooky.htm
Recognized record 1
REGION_ID: 0
PRICE: 20 367,- or 17 115,- or 17 609,-
PRODUCER: ASUS
DATE: 2007-11-21 14:38:19.289
SIZE_OF_HDD: 160GB
TOKENIZED_TEXT: Inovativní zajímavý notebook nabízející dokonalou mobilitu|maximální výkon|uspokojivou výdrž na ... |ASUS F3KE-AP012|A64 X2 TK53|15.4" WXGA CS|1GB|X2300 128MB|160GB 5.4k|DVD±RW|WF|BT|CAM|Bez OS|Skladem 3-5 ks|s DPH 20 367,-|17 115,-|17 609,-|-2,8% |NA426l|kód:|
SIZE_OF_RAM: 1GB or 128MB
LINKS_TO_DETAIL: /notebook-asus-f3ke-ap012-d79797.htm
DISPLAY_SIZE: 15.4"
Recognized record 2
REGION_ID: 0
PRICE: 21 190,- or 17 807,- or 20 796,-
PRODUCER: MSI
DATE: 2007-11-21 14:38:19.289
SIZE_OF_HDD: 120GB
TOKENIZED_TEXT: Univerzální notebook v elegantním designu|který každého zaujme na první pohled. Nabízí špičkový ... |MSI MegaBook L730X-074CZ|T64 X2 TL52|17" WXGA|1GB|120GB 5.4k|DVD±RW+DL|WF|BT|bez OS|Skladem 1-2 ks|s DPH 21 190, |17 807, |20 796, |-14,4% |NB024e|kód:|
SIZE_OF_RAM: 1GB
LINKS_TO_DETAIL: /notebook-msi-megabook-l730x-074cz-d74966.htm
DISPLAY_SIZE: 17"
Recognized record 3
REGION_ID: 0
WiFi: WiFi
PRICE: 14 744,- or 12 390,- or 14 559,-
PRODUCER: ACER
DATE: 2007-11-21 14:38:19.289
SIZE_OF_HDD: 80GB
TOKENIZED_TEXT: Zajímavý inovativní notebook založený na procesoru Mobile AMD Sempron s technologií AMD PowerNow! ... |ACER Aspire 3105WLMi|S64 3600+|15.4" WXGA CB|1GB|80GB 5.4k|DVD±RW|WiFi|VIS HB|Skladem 1-2 ks|s DPH 14 744,-|12 390,-|14 559,-|-14,9% |NC205n|kód:|
SIZE_OF_RAM: 1GB
LINKS_TO_DETAIL: /notebook-acer-aspire-3105wlmi-d80383.htm
DISPLAY_SIZE: 15.4"
Recognized record 4
REGION_ID: 0
PRICE: 17 016,- or 14 299,- or 15 869,-
PRODUCER: ACER
DATE: 2007-11-21 14:38:19.289
Find: opel
Next
Previous
Highlight all
Match case
Done

# Anotace textových zdrojů

Diplomová práce

Jan Dědek

# Techniky UFAL použité pro web content mining

HZS Jihomoravského kraje

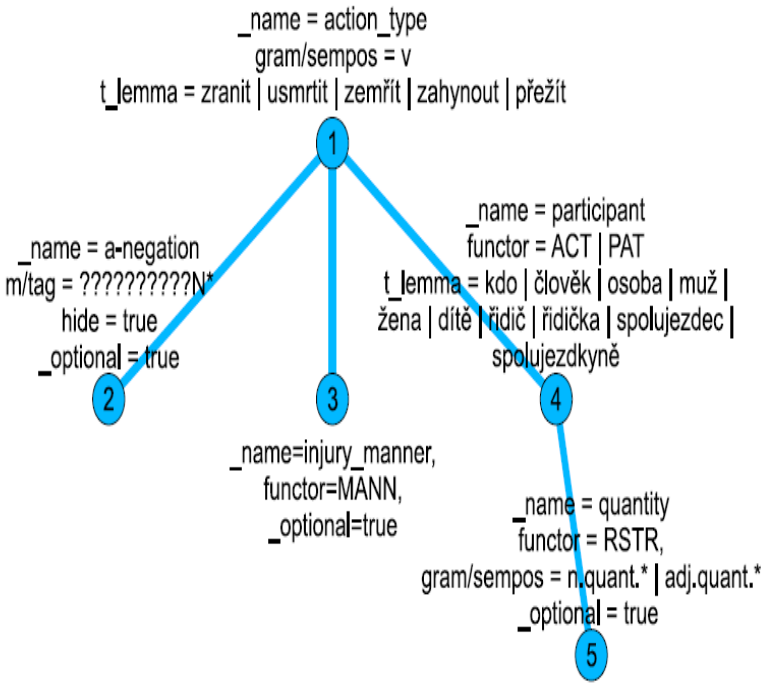
Zubatého 1, 614 00 Brno, telefon 950 630 111,
http://www.firebrno.cz
Zpravodajství v roce 2006

06.09.2007

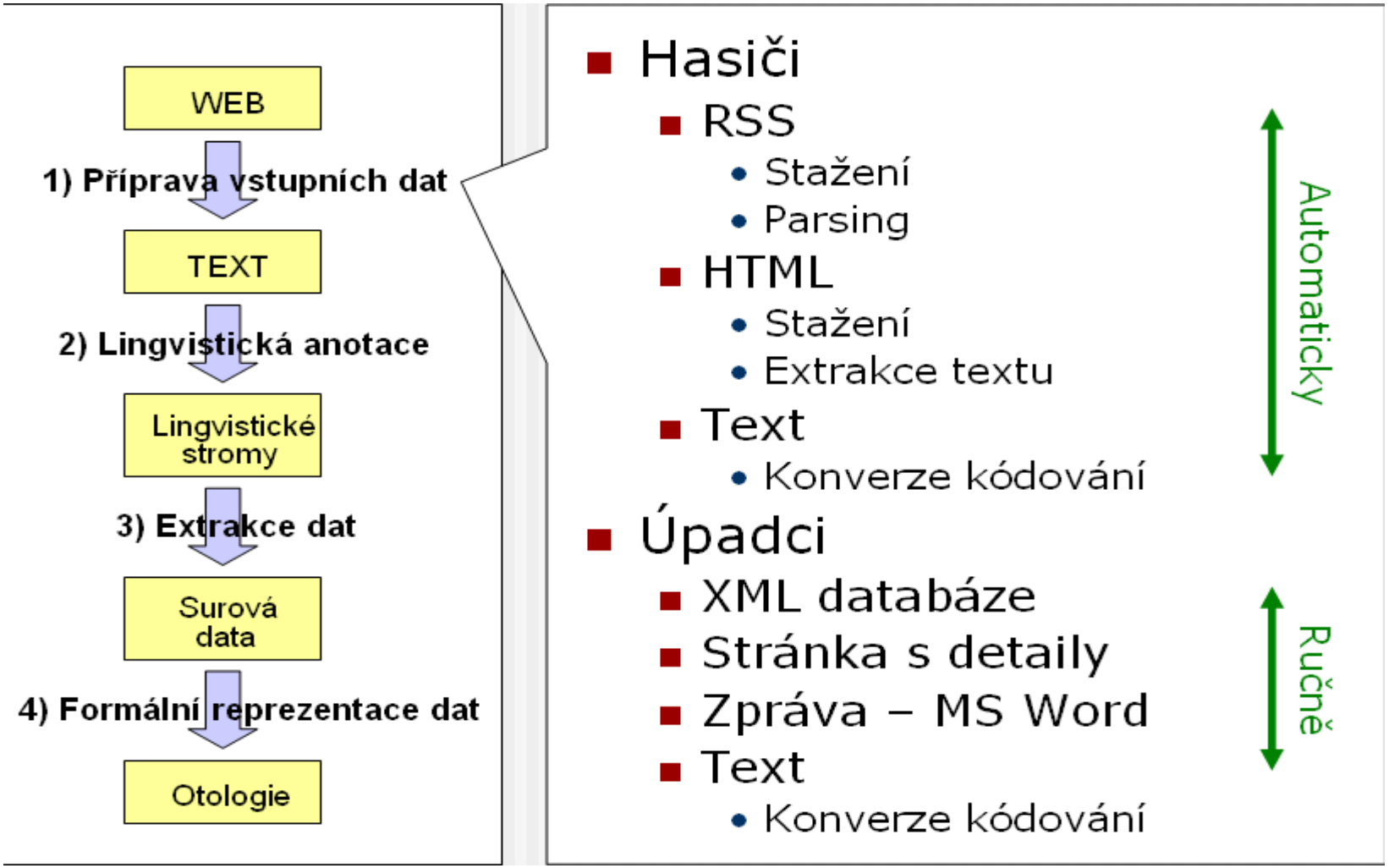
## Při nehodě zemřel řidič

***Hasiči v Jihomoravském kraji dnes zatím vyjížděli k sedmi dopravním nehodám. Nehody si vyžádaly jeden lidský život a několik zranění.***

Šestašedesátiletý řidič vozidla Opel Vectra zemřel při nehodě u obce Zbraslav na Brněnsku. Muž vyjel do protisměru, kde se střetl s vozidlem Škoda Fabia. V kabině zdemolovaného Opelu zůstal řidič zaklíněn za nohy a přivolaný lékař konstatoval, že podlehl zraněním. Padesátiletý řidič Škody Fabia, která po střetu skončila zhruba dva metry od silnice v poli, utrpěl zranění.



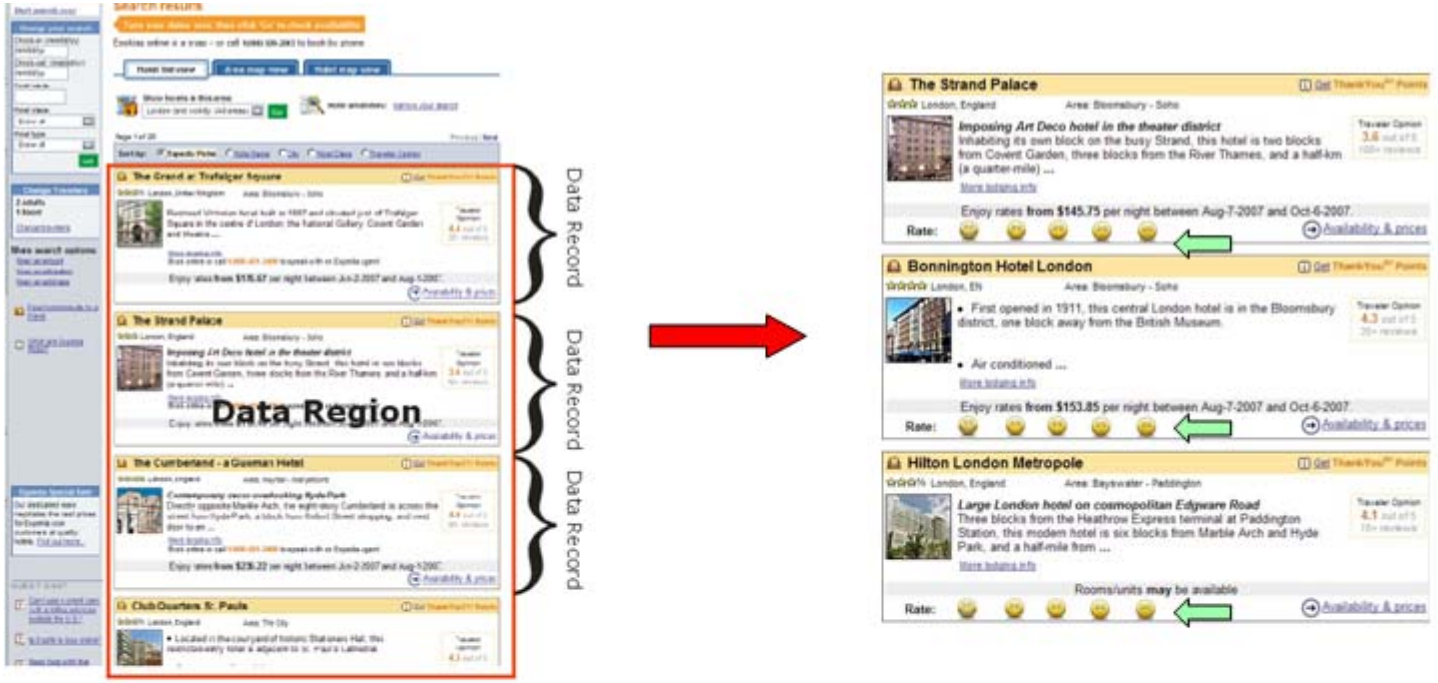
# Řešení s použitím UFAL-ware



# Modely uživatelských preferencí

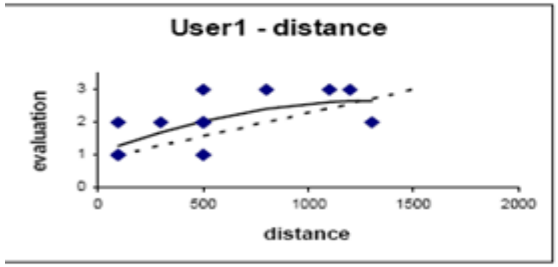
Alan Eckhardt, Tomáš Horváth,
Jaroslav Pokorný, Peter Vojtáš

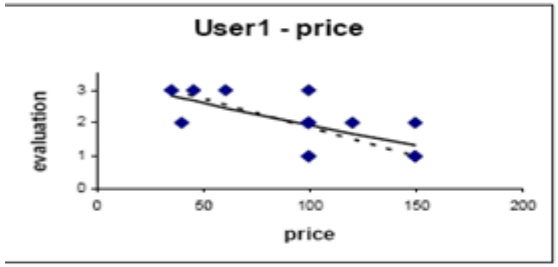
# Uživatel ohodnotí vzorek

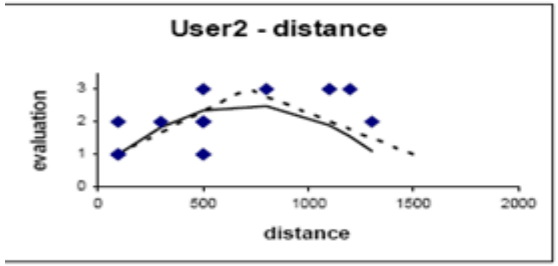


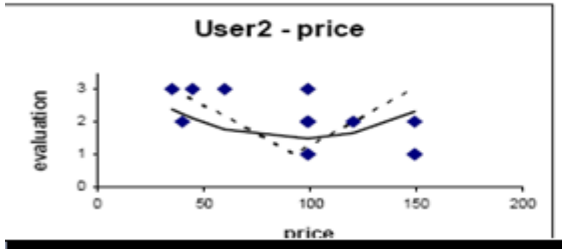

| Hotels properties | | | | Users evaluation | | | |
|---|---|---|---|---|---|---|---|
| | | | | User1 | | User 2 | |
| Name | Distance | Price | Equipment | grade | num. | grade | num. |
| Apple | 100 m | 99 $ | nothing | poor | 1 | poor | 1 |
| Danube | 1300 m | 120 $ | tv | good | 2 | poor | 1 |
| Cherry | 500 m | 99 $ | Internet | good | 2 | good | 2 |
| Iris | 1100 m | 35 $ | internet, tv | excellent | 3 | excellent | 3 |
| Lemon | 500 m | 149 $ | nothing | poor | 1 | excellent | 3 |
| Linden | 1200 m | 60 $ | internet, tv | excellent | 3 | poor | 1 |

# Model



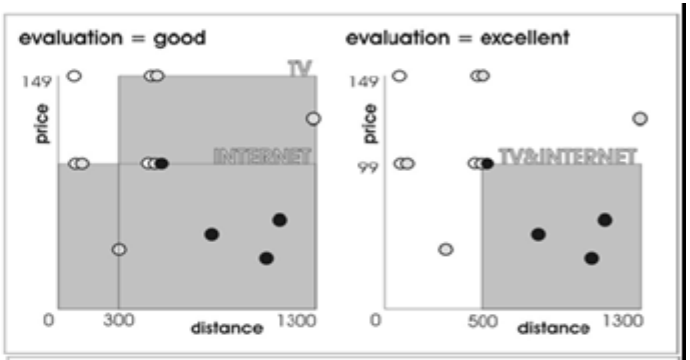



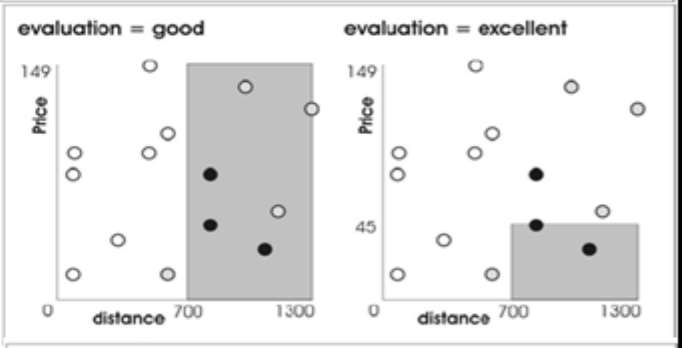



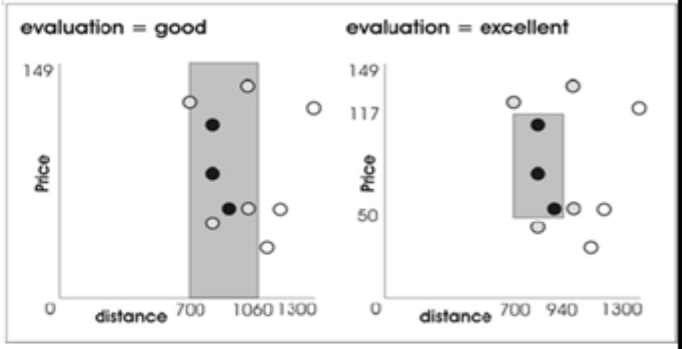




Učení lokálních preferencí

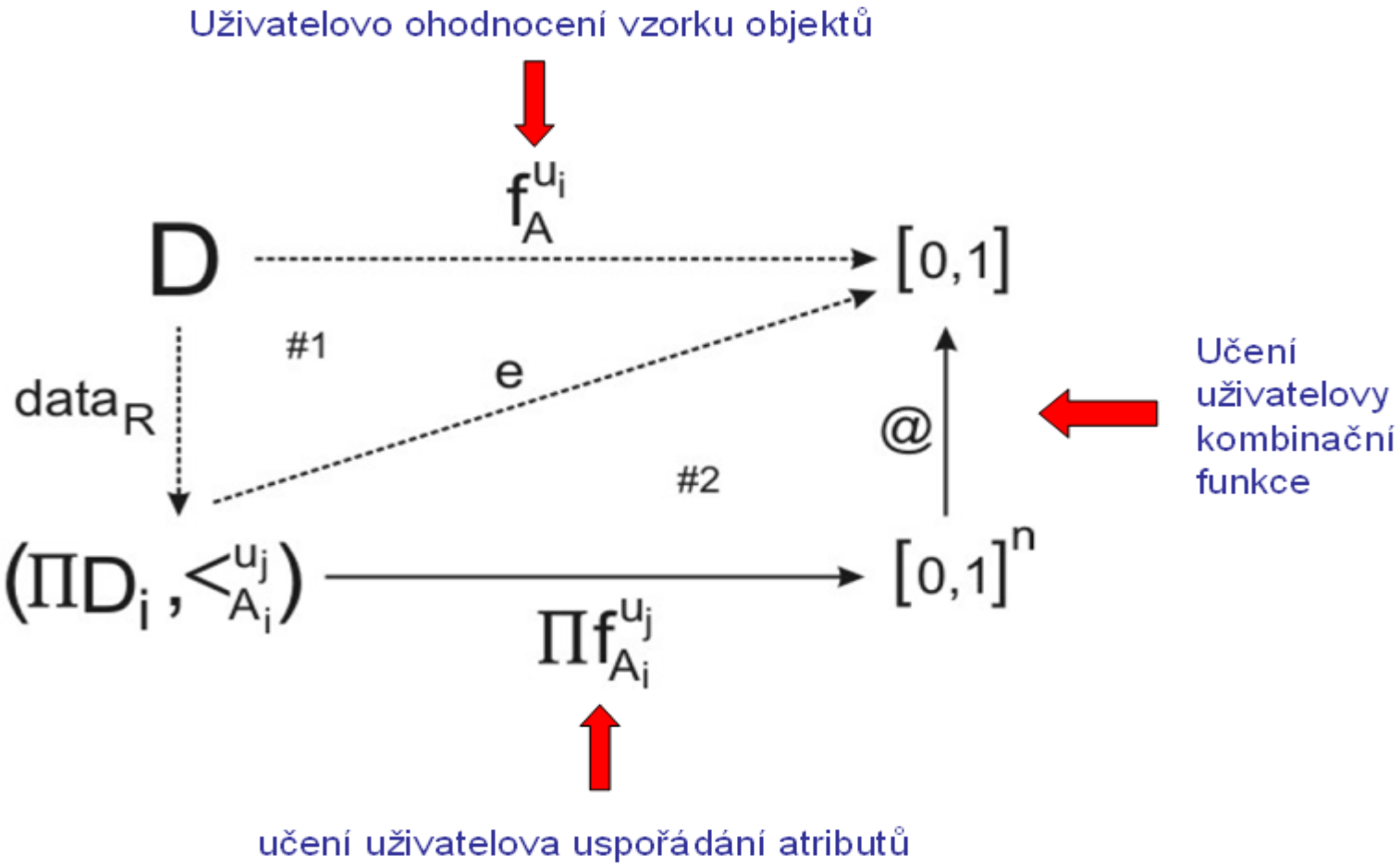
Učení globální kombinační funkce iterací



# Model


Uživatelovo ohodnocení vzorku objektů
$f_A^{u_i}$
D
[0,1]
#1
e
$data_R$
@
Učení uživatelovy kombinační funkce
#2
$(\Pi D_i, <_{A_i}^{u_j})$
$[0,1]^n$
$\Pi f_{A_i}^{u_j}$
učení uživatelova uspořádání atributů

# Top-k dotazů TOKAF a víceuživatelské indexy

Diplomová práce Alan Eckhardt

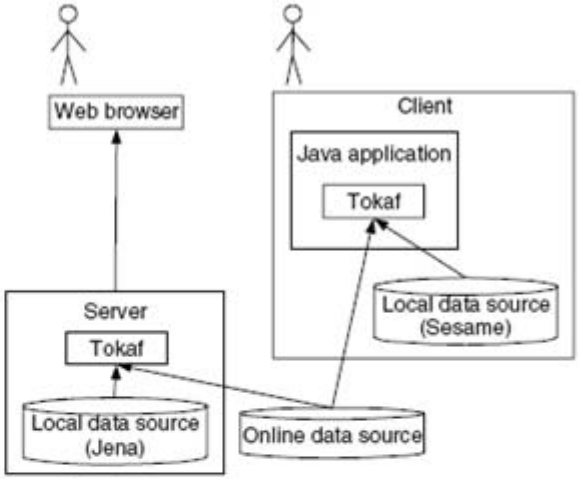
+

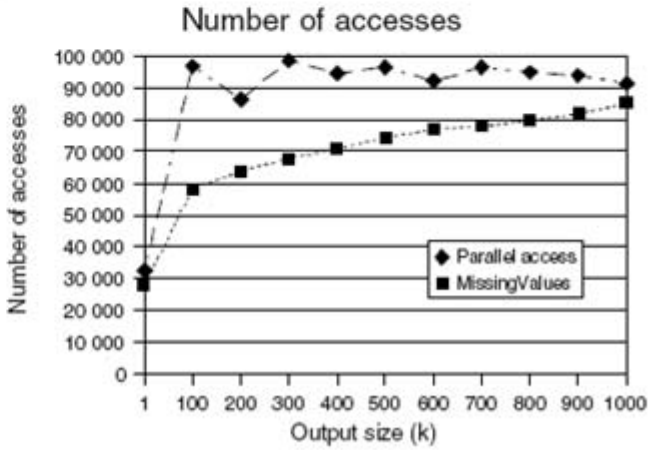
AE, JP, PV

# Implementation

# Experiments

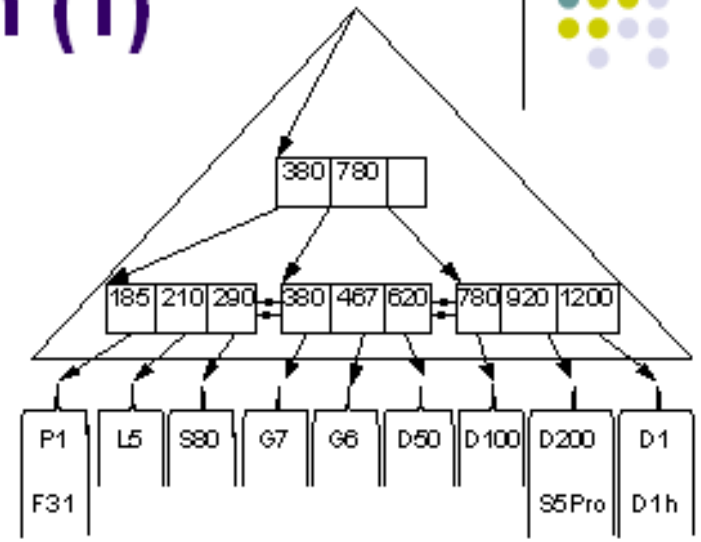
- Fagin's top-k
- TOKAF
- IGAP-ALEPH

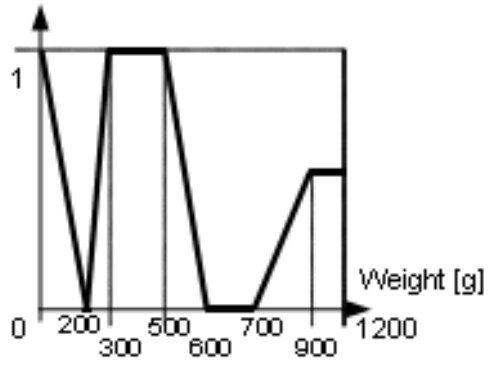


# Using indexes for ordering of an attribute domain (1)

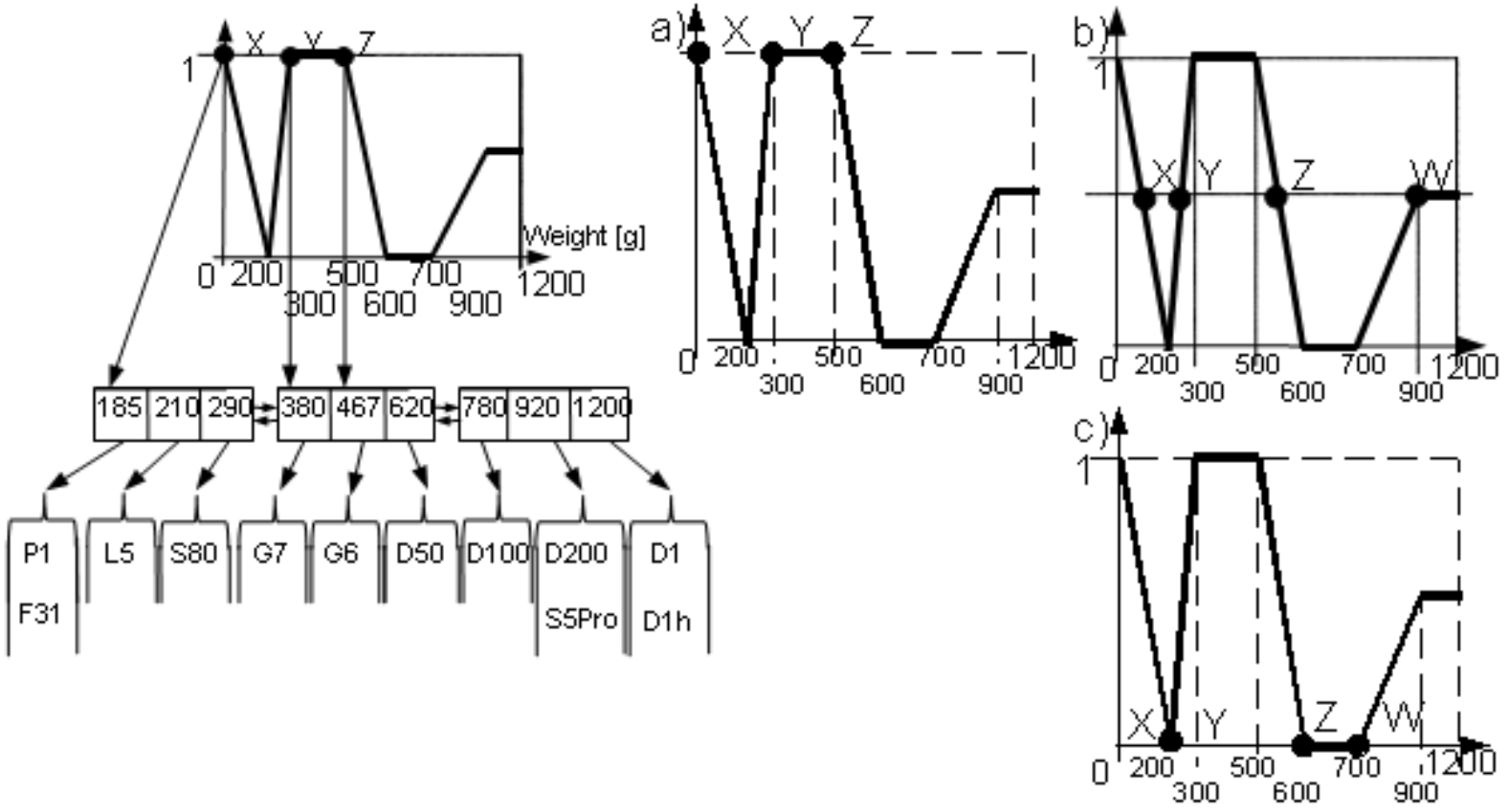
- For top-k search, attribute domains have to be ordered
  - According to a fuzzy function
- Using an index for numerical domains
- Limiting fuzzy functions
  - Intervals of monotonicity
  - For easier representation



# Using indexes for ordering of an attribute domain (2)

- Using pointers to the index for ordering



# Otevřené problémy

- Integrace nástrojů
- Uživatelské experimenty
- Přenesitelnost
- Obtížnost zdrojů
- Množství lidské práce
- Učení ontologie
- Automatizace
- Velikost dat, čas odezvy
- Přesnost, úplnost, …

Děkuji za pozornost