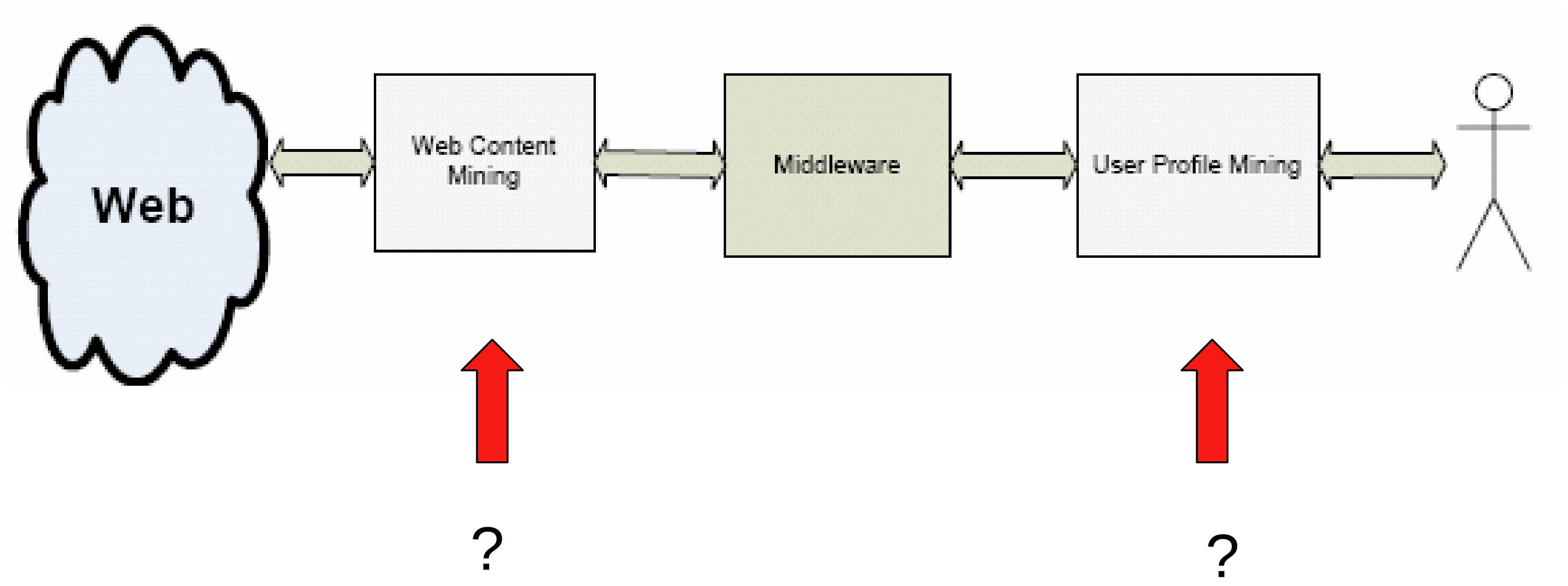
Uncertainty Issues in Automating Process Connecting Web and User

A. Eckhardt, T. Horváth, D. Maruščák,*
R. Novotný, P. Vojtáš*

Charles University Prague
* P. J. Safarik University Kosice

Motivation

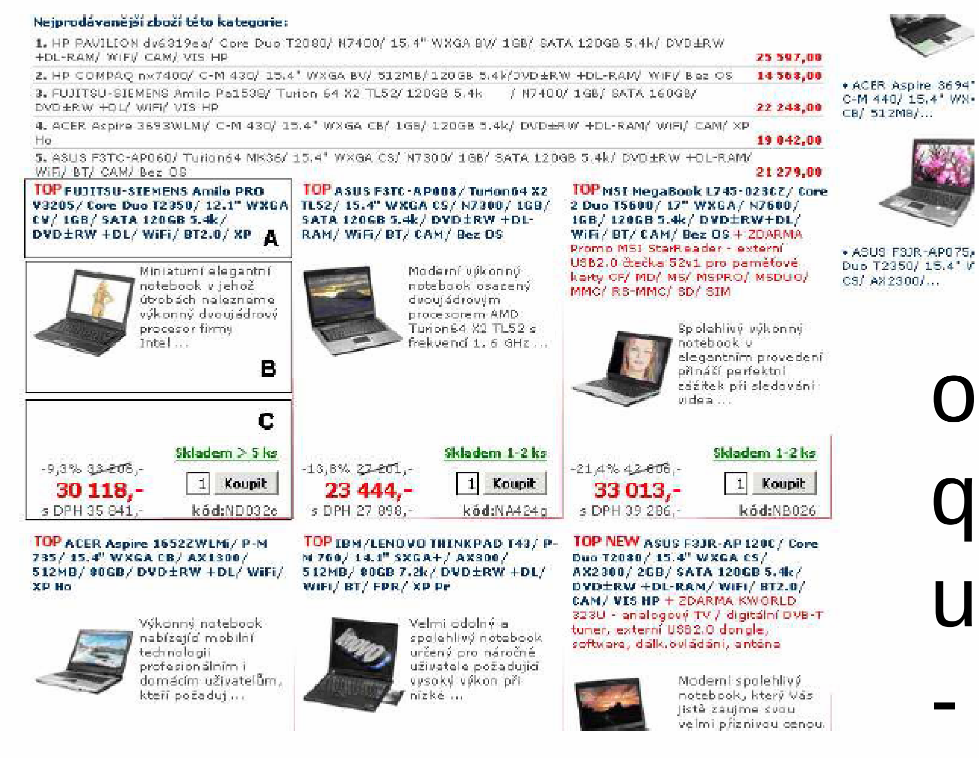


Users looking for a hotel, notebook, car,… on the web without semantic labels

Automating the process brings uncertainty

Motivation

Human understandable not machine readable



Nejprodávanější zboží této kategorie:

1. HP PAVILION dv6319ea/ 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

• ACER Aspire 3694" C-M 440/ 15.4" WX• CB/ 512MB/...

• ASUS F3JR-AP075/ Duo T2350/ 15.4" V CS/ AX2300/...

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 nalezneme výkonný dvoujádrový procesor firmy Intel ...

B

C

Skladem > 5 ks

-9,3% ~~33 206~~,-

30 118,-

1 Koupit

s DPH 35 841,-

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,-

1 Koupit

s DPH 27 898,-

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% ~~42 006~~,-

33 013,-

1 Koupit

s DPH 39 286,-

kód:NB026

TOP ACER Aspire 1652ZWLMi/ 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.

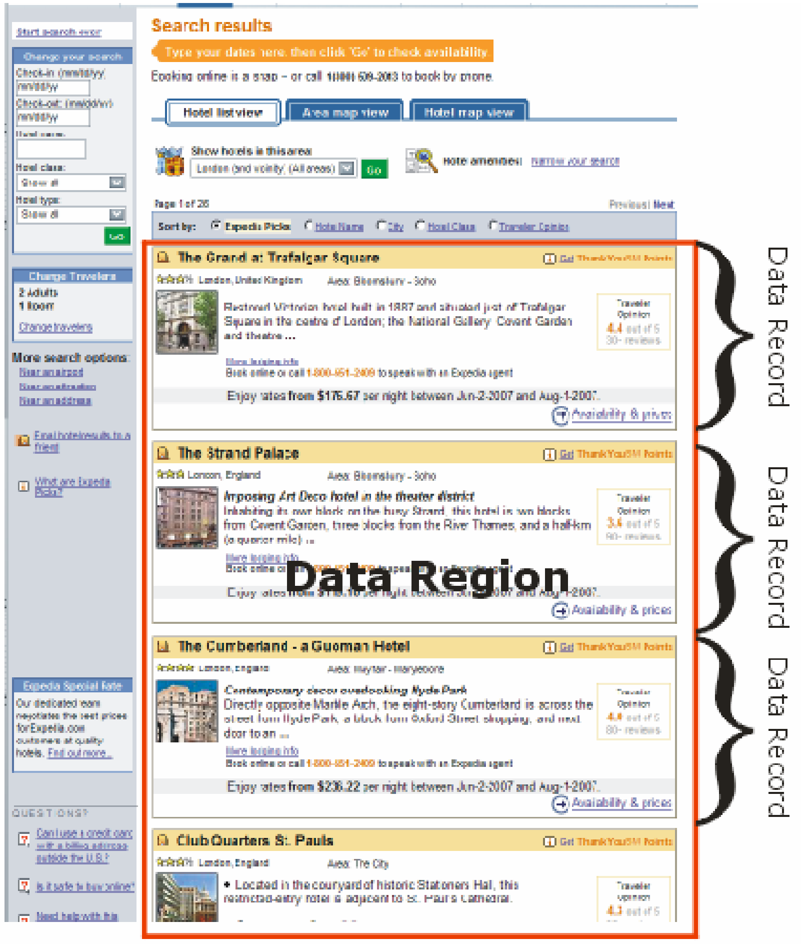
only conjunctive user queries – no individual user preferences –
- boring click through

Outline

- Motivation
- Uncertainty in web content mining
  - discovering data regions, data records
  - attribute values extraction
- Middleware
- Uncertainty in user preference mining
  - learning attribute preference
  - learning combination function
- Experiments – uncertainty issues detected
- Conclusions

Uncertainty in web content mining

- Crawling web (not here)
- Discovering relevant pages (not here)
- Discovering data regions
- Discovering data records



Start search over

Change your search

Check-in (mm/dd/yy)

mm/dd/yy

Check-out (mm/dd/yy)

mm/dd/yy

Hotel name:

Hotel class: Show all

Hotel type: Show all

Go

Change Travelers

2 Adults
1 Room

Change travelers

More search options:

Near an airport
Near an attraction
Near an address

Email hotel results to a friend

What are Expedia Picks?

Expedia Special Rate

Our dedicated team negotiates the best prices for Expedia.com customers at quality hotels. Find out more...

QUESTIONS?

Can I use a credit card with a billing address outside the U.S.?

Is it safe to buy online?

Need help with this

Search results

Type your dates here, then click 'Go' to check availability

Booking online is a snap – or call 1(800) 509-2083 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: Narrow your search

Page 1 of 26

Previous | Next

Sort by: ● Expedia Picks ○ Hotel Name ○ City ○ Hotel Class ○ Traveler Opinion

The Grand at Trafalgar Square — Get ThankYou℠ Points

London, United Kingdom — Area: Bloomsbury - Soho

Restored Victorian hotel built in 1887 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

More lodging info

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 — Get ThankYou℠ Points

London, England — Area: Bloomsbury - Soho

Imposing Art Deco hotel in the theater district

Inhabiting its own block on the busy Strand, this hotel is two blocks from Covent Garden, three blocks from the River Thames, and a half-km (a quarter mile) ...

Traveler Opinion 3.6 out of 5, 80+ reviews

More lodging info

Availability & prices

The Cumberland - a Guoman Hotel — Get ThankYou℠ Points

London, England — Area: Mayfair - Marylebone

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 ...

Traveler Opinion 4.4 out of 5, 80+ reviews

More lodging info

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 — Get ThankYou℠ Points

London, England — Area: The City

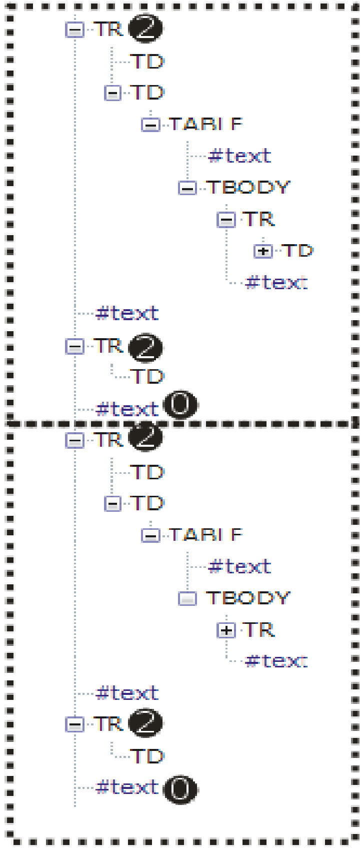
- Located in the courtyard of historic Stationers Hall, this restricted-entry hotel is adjacent to St. Paul's Cathedral.

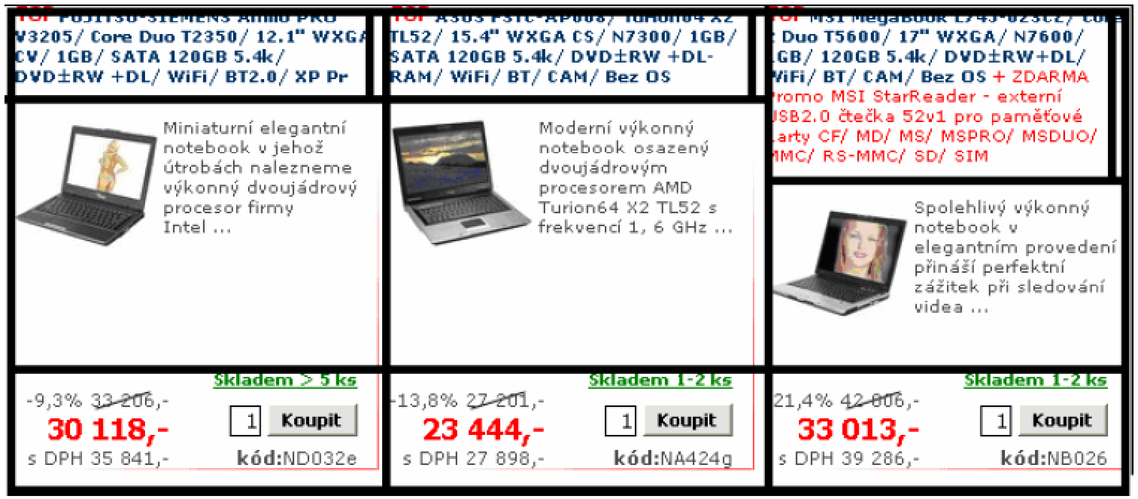
Traveler Opinion 4.3 out of 5

Uncertainty in web content mining

Search over DOM form of page

Non-contiguous records





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 Pr

Miniaturní elegantní notebook v jehož útrobách nalezneme výkonný dvoujádrový procesor firmy Intel ...

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 L745-025CZ/ Core 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% ~~42 006~~,-

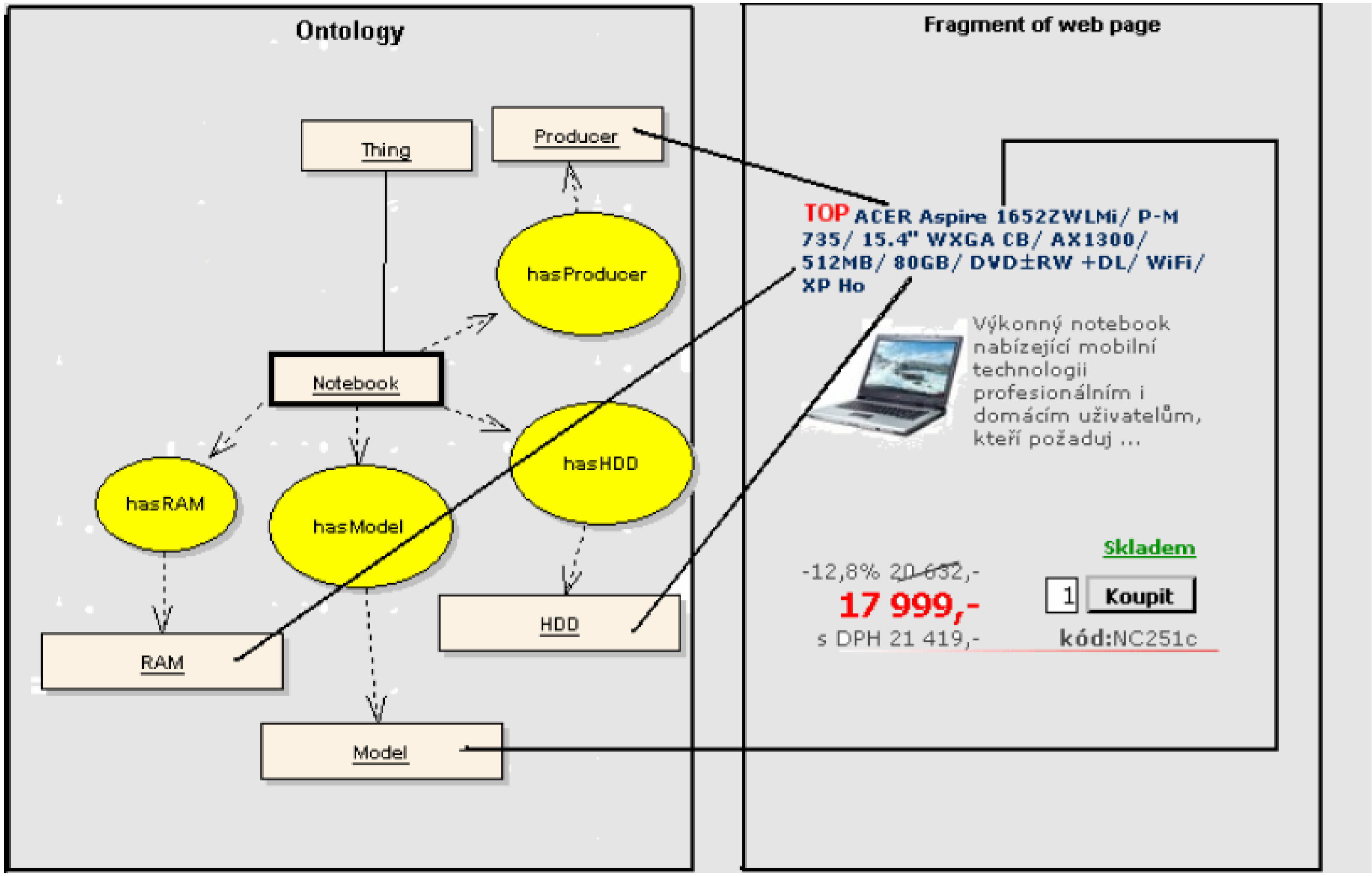
33 013,-

s DPH 39 286,-

1 Koupit

kód:NB026

Ontology based attribute value extraction



TOP ACER Aspire 1652ZWLMi/ 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 ...

Skladem

-12,8% ~~20 632,-~~

17 999,-

s DPH 21 419,-

1 Koupit

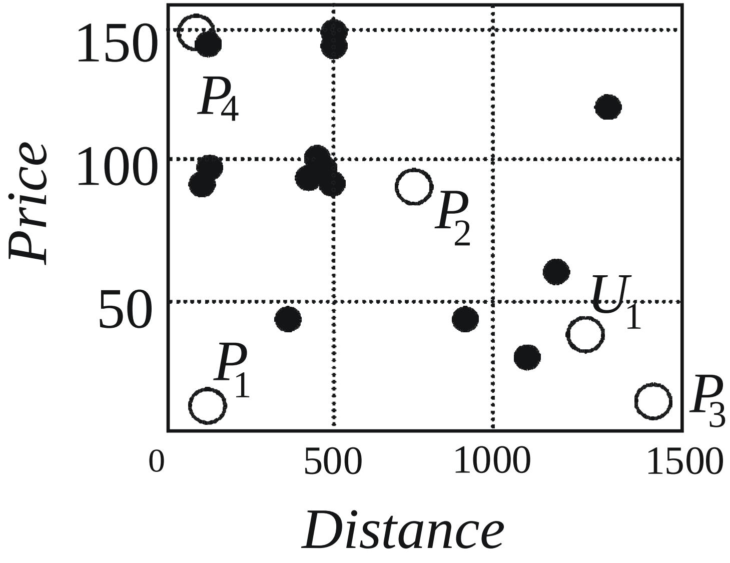
kód:NC251c

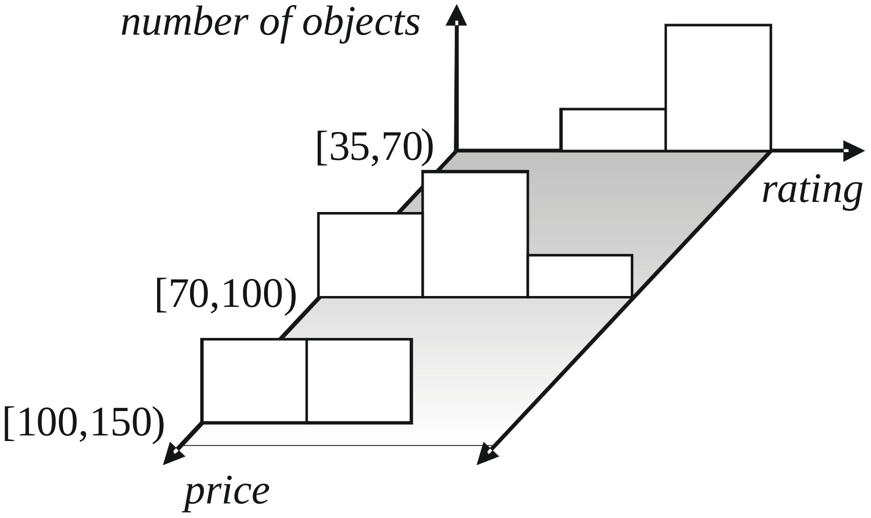
Additional low level extraction ontology

```
<owl:DatatypeProperty rdf:ID="hasPrice">
 <rdfs:domain rdf:resource="#Hotel"/>
 <p1:maxLength
     rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
     10
 </p1:maxLength>
<p1:pattern
     rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
   (\$)? ?[\d]{1,10} ?(.){1,3}
 </p1:pattern>
  <rdfs:label
     rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
    PRICE
  </rdfs:label>
</owl:DatatypeProperty>
```

Middleware, top-k, user model

- Storing data extracted from web pages
- Supporting top-k queries based on user’s combination of user’s attribute preferences





Uncertainty in user preference mining

- Detecting attribute preferences – cheap or expensive, close or far…domain dependent
- Combination of attribute preferences helps order incomparable objects – IGAP method

User1_hotel(H) good in degree at least @(f_1(x), f_2(y), …)

IF User1_hotel_price(x) good in degree at least f_1(x) **AND**

User1_hotel_distance(y) good in degree at least f_2(y)

Implementation, experiments

- Modular implementation which allows additional modules to be incorporated (e. g. querying with preference-based querying)
- Communication between modules is based on the traditional Observer/Listener design pattern
- Middleware system for performing top-k queries over RDF data
- As a Java library, our system can be used either on the server side, for example in a Web service, or on the client side
- General method using B+ trees to simulate arbitrary fuzzy ordering of a domain

Identified uncertainty issues

- identifying HTML nodes with relevant information in the sub-tree,
- tuning similarity measures for discovery of similar tag subtrees,
- identifying single data records in non-contiguous html source,
- extracting attribute values
- learning user’s preferences of particular attributes
- learn the user preference combination function.

Conclusions

- Automating the process of user dependent web search – causes uncertainty
- Identified uncertainty issues in our approach (other approaches may have other uncertainty challenges)
- Whole process, querying, results are uncertain – creating a web service we need UIF – Uncertainty Interchange Format

Conclusions

- Implementation and experiments in different domains

  some domains are “easier” to mine – e.g. notebooks – results are more certain

  some domains are “more difficult” – e.g. hotels – results are more uncertain
- Human training time, learning ontology,…
- Low level extraction ontology
- Future work

Thank you

Questions?