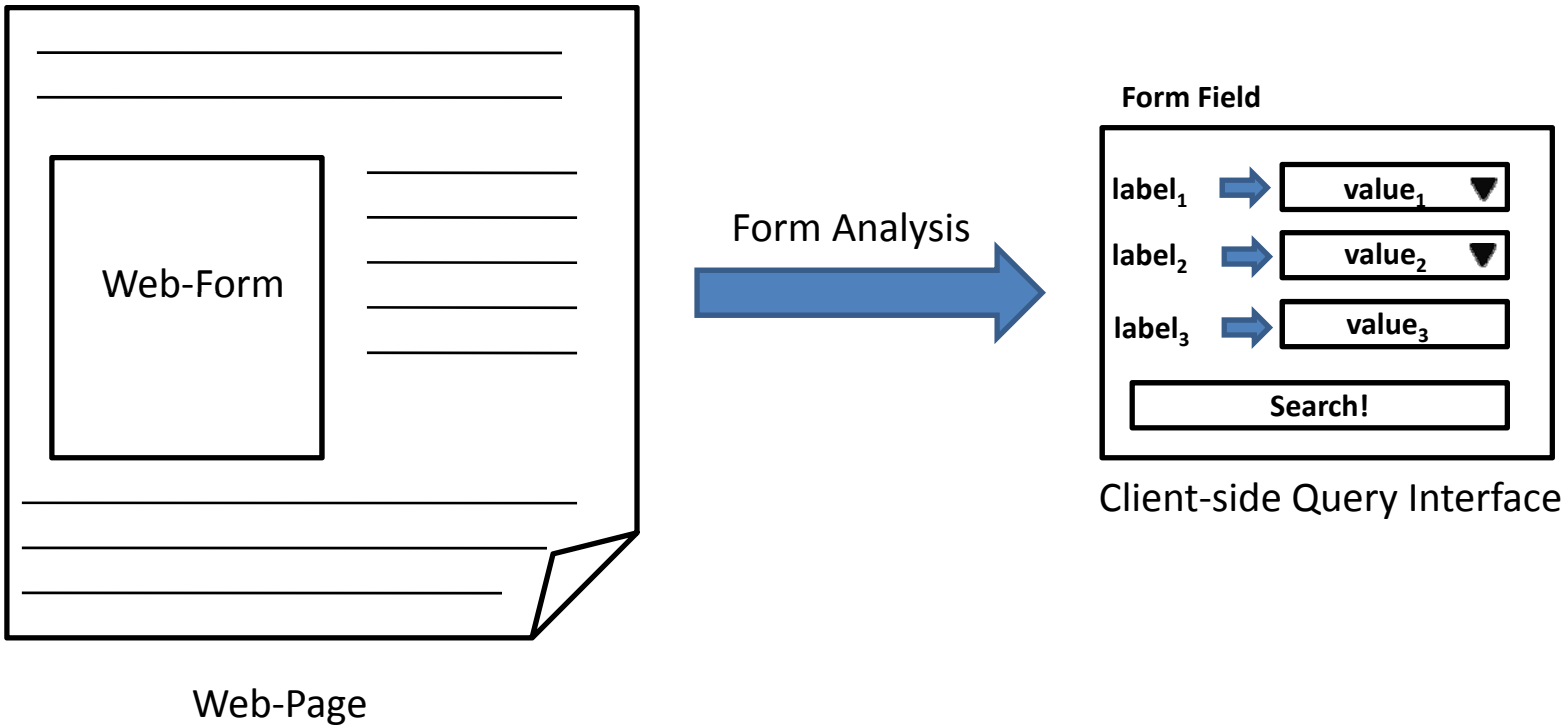
Deep Web Navigation by Example

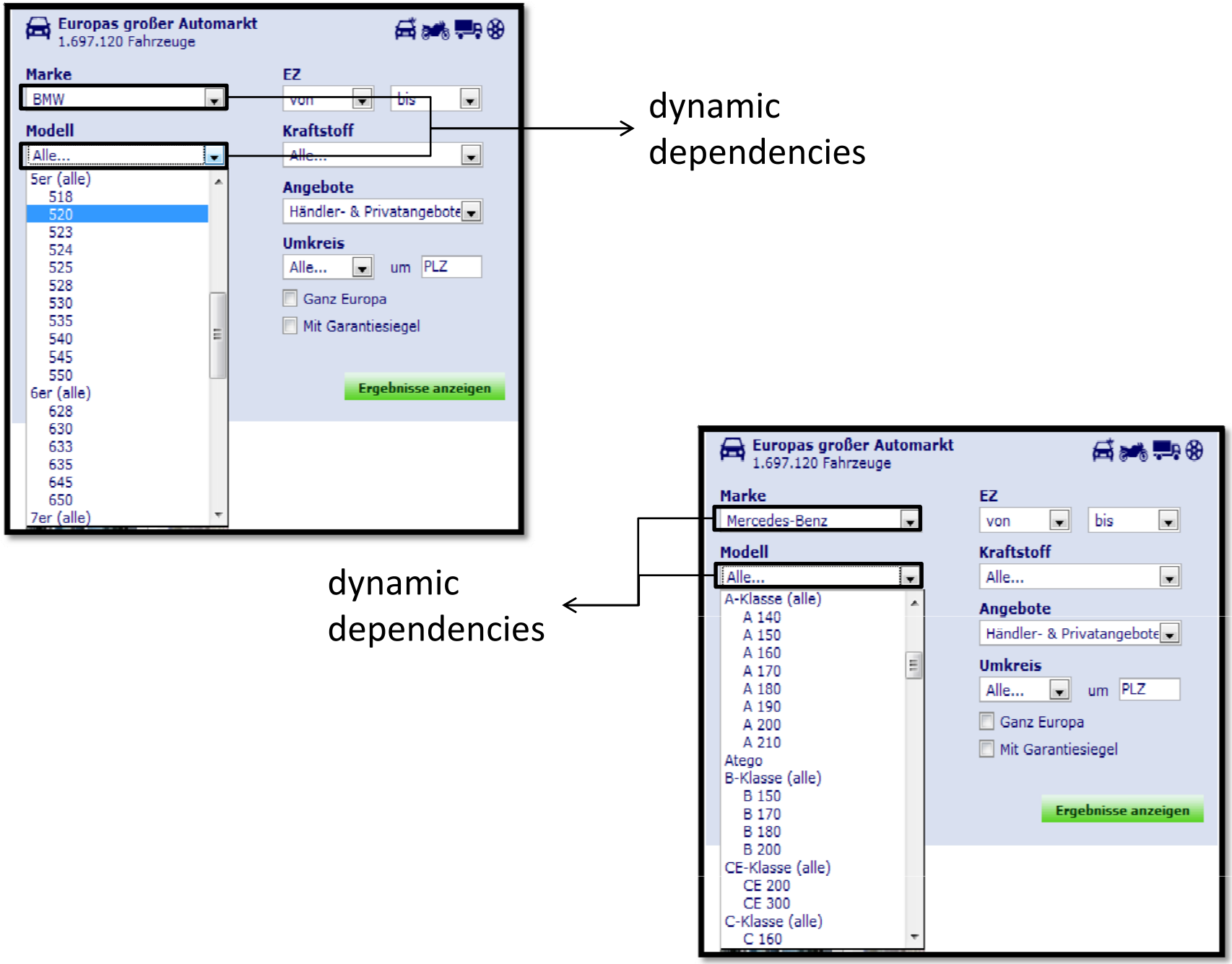
Yang Wang, Thomas Hornung

Motivation

- Transformation of Web forms into machine-accessible Query Interfaces
- Problems solved by this work:
  - Form Analysis (Part 1)
    - Dynamic dependencies of input elements
  - Deep Web Navigation to result page (Part 2)

1 Form Analysis

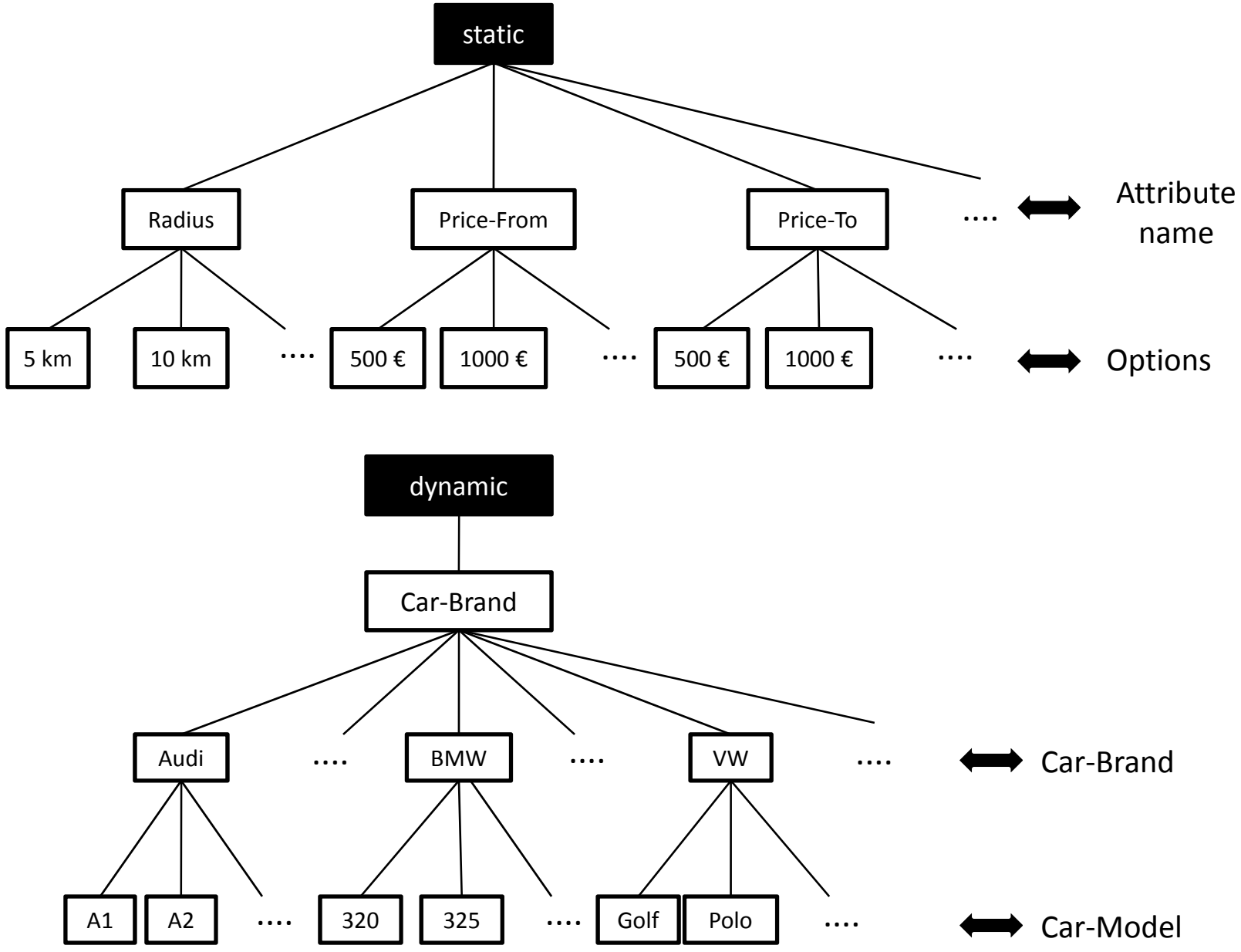

Web-Form
Form Analysis
Form Field
label1
value1
label2
value2
label3
value3
Search!
Client-side Query Interface
Web-Page


Europas großer Automarkt
1.697.120 Fahrzeuge
Marke
BMW
Modell
Alle...
5er (alle)
518
520
523
524
525
528
530
535
540
545
550
6er (alle)
628
630
633
635
645
650
7er (alle)
EZ
von
bis
Kraftstoff
Alle...
Angebote
Händler- & Privatangebote
Umkreis
Alle...
um
PLZ
Ganz Europa
Mit Garantiesiegel
Ergebnisse anzeigen
dynamic dependencies
dynamic dependencies
Europas großer Automarkt
1.697.120 Fahrzeuge
Marke
Mercedes-Benz
Modell
Alle...
A-Klasse (alle)
A 140
A 150
A 160
A 170
A 180
A 190
A 200
A 210
Atego
B-Klasse (alle)
B 150
B 170
B 180
B 200
CE-Klasse (alle)
CE 200
CE 300
C-Klasse (alle)
C 160
EZ
von
bis
Kraftstoff
Alle...
Angebote
Händler- & Privatangebote
Umkreis
Alle...
um
PLZ
Ganz Europa
Mit Garantiesiegel
Ergebnisse anzeigen

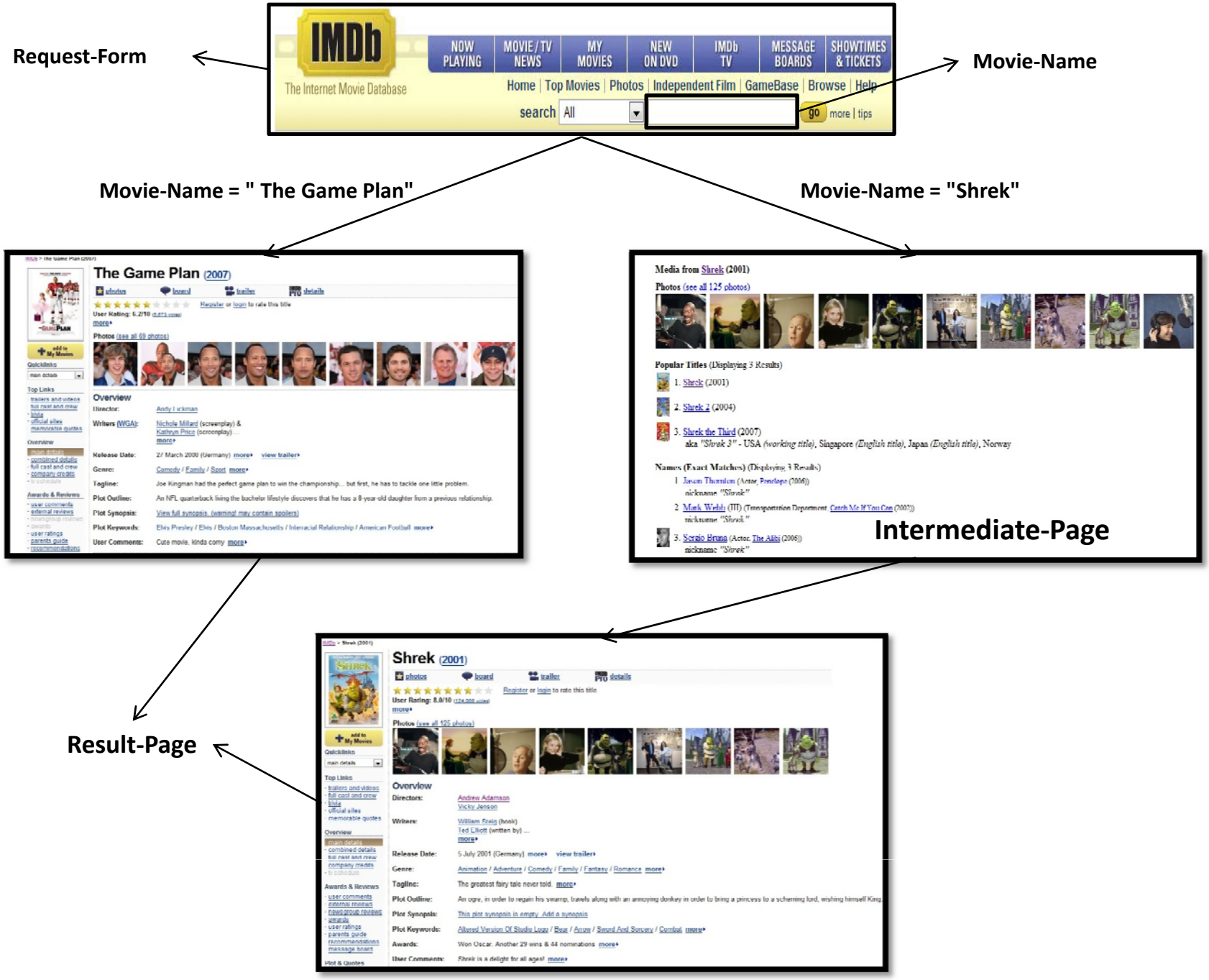
Dynamic or Static?

Idea: Simulate HTML-events

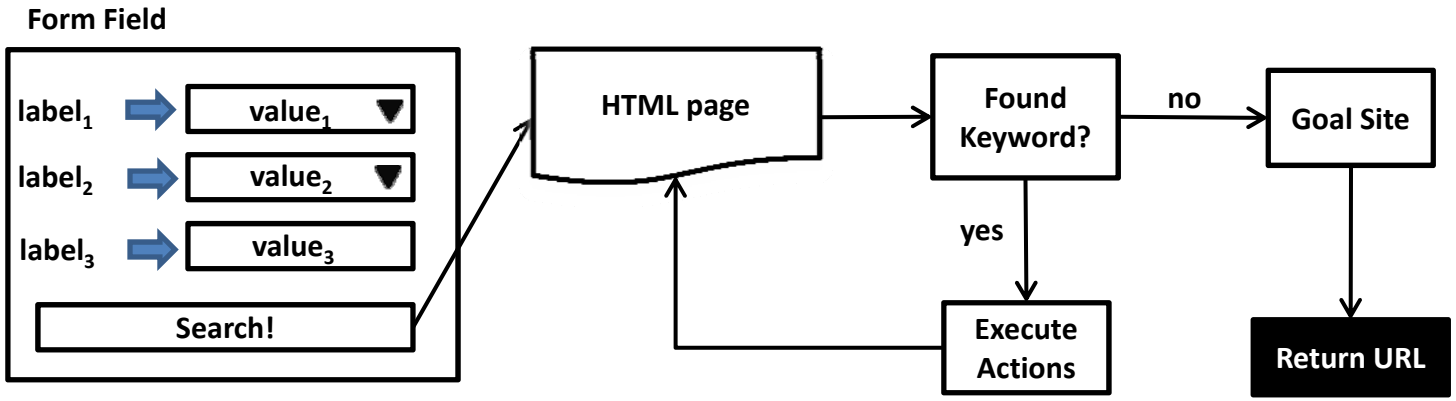
1. Save initial status of marked dropdown menus (temporally)
   a. Length of options
   b. Option-text
2. Change selected option of one marked dropdown menu
3. Check the options of other marked dropdown menu
   a. changed → dynamic
   b. unchanged → static
4. Text-Input-Field → static

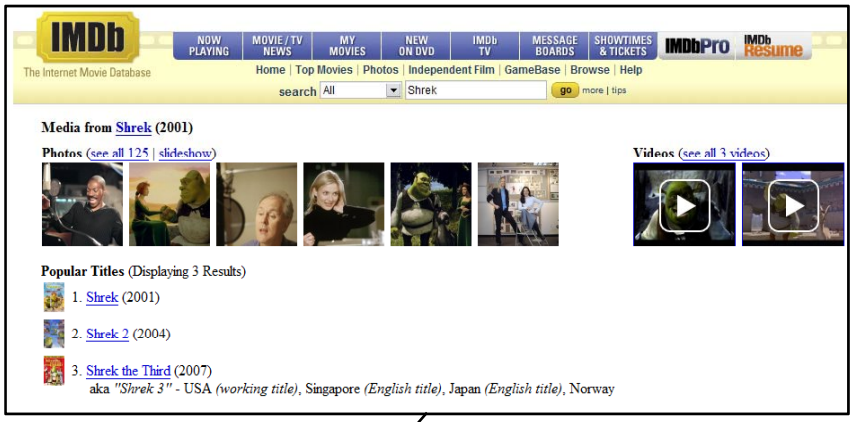

static
Radius
Price-From
Price-To
....
Attribute name
5 km
10 km
....
500 €
1000 €
....
500 €
1000 €
....
Options
dynamic
Car-Brand
Audi
....
BMW
....
VW
....
Car-Brand
A1
A2
....
320
325
....
Golf
Polo
....
Car-Model

2 Deep Web Navigation


Request-Form
IMDb
The Internet Movie Database
NOW PLAYING
MOVIE / TV NEWS
MY MOVIES
NEW ON DVD
IMDb TV
MESSAGE BOARDS
SHOWTIMES & TICKETS
Home | Top Movies | Photos | Independent Film | GameBase | Browse | Help
search All
go more | tips
Movie-Name
Movie-Name = " The Game Plan"
Movie-Name = "Shrek"
The Game Plan (2007)
Overview
Director:
Writers (WGA):
Release Date:
Genre:
Tagline:
Plot Outline:
Plot Synopsis:
Plot Keywords:
User Comments:
Media from Shrek (2001)
Photos (see all 125 photos)
Popular Titles (Displaying 3 Results)
1. Shrek (2001)
2. Shrek 2 (2004)
3. Shrek the Third (2007)
Names (Exact Matches) (Displaying 3 Results)
Intermediate-Page
Result-Page
Shrek (2001)
Overview
Directors:
Writers:
Release Date:
Genre:
Tagline:
Plot Outline:
Plot Synopsis:
Plot Keywords:
Awards:
User Comments:

Navigation Process



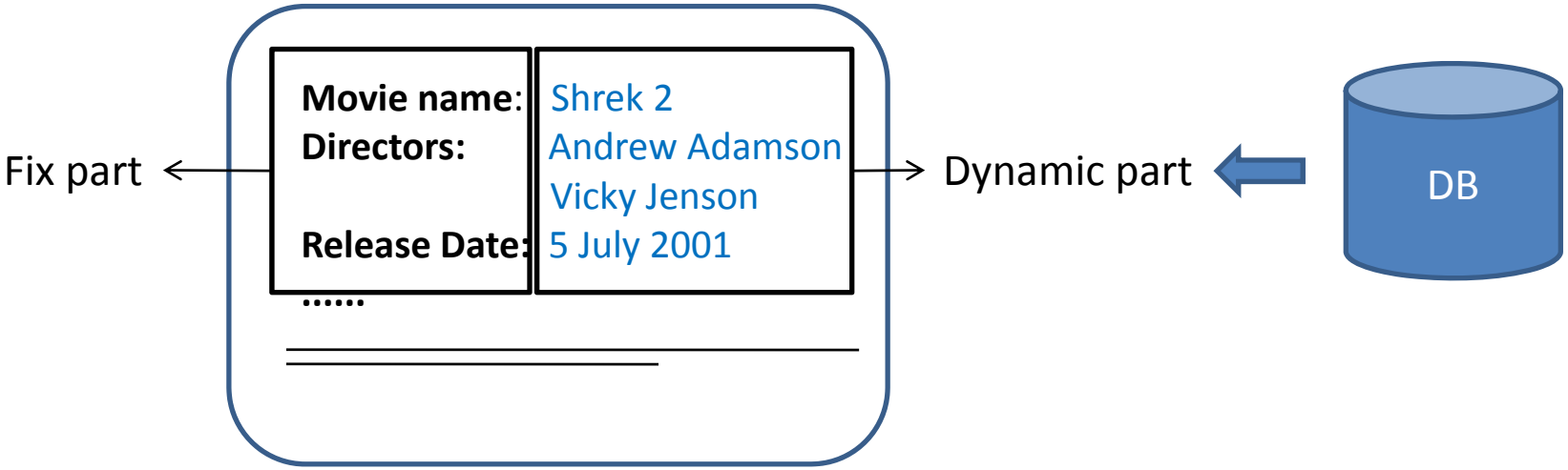

IMDb
The Internet Movie Database
NOW PLAYING
MOVIE / TV NEWS
MY MOVIES
NEW ON DVD
IMDb TV
MESSAGE BOARDS
SHOWTIMES & TICKETS
IMDbPro
IMDb Resume
Home | Top Movies | Photos | Independent Film | GameBase | Browse | Help
search All Shrek go more | tips
Media from Shrek (2001)
Photos (see all 125 | slideshow)
Videos (see all 3 videos)
Popular Titles (Displaying 3 Results)
1. Shrek (2001)
2. Shrek 2 (2004)
3. Shrek the Third (2007)
aka "Shrek 3" - USA (working title), Singapore (English title), Japan (English title), Norway


IMDb
The Internet Movie Database
NOW PLAYING
MOVIE / TV NEWS
MY MOVIES
NEW ON DVD
IMDb TV
MESSAGE BOARDS
SHOWTIMES & TICKETS
IMDbPro
IMDb Resume
Home | Top Movies | Photos | Independent Film | GameBase | Browse | Help
search All Spiderman go more | tips
Media from Spider-Man (2002)
Photos (see all 105 | slideshow)
Videos (see all 3 videos)
Popular Titles (Displaying 3 Results)
1. Spider-Man (2002)
aka "Spider-Man: The Motion Picture" - USA (working title)
aka "Spiderman" - USA (alternative spelling)
2. Spider-Man 2 (2004)
aka "Spider-Man 2.1" - USA (recut version)
aka "The Amazing Spider-Man" - USA (working title)
aka "Spider-Man 2: The IMAX Experience" - USA (IMAX version)
aka "Spider-Man: No More" - USA (working title)
aka "Spiderman 2" - USA (alternative spelling)
aka "Spider-Man 2 Lives" - USA (working title)

similar


IMDb
The Internet Movie Database
NOW PLAYING
MOVIE / TV NEWS
MY MOVIES
NEW ON DVD
IMDb TV
MESSAGE BOARDS
SHOWTIMES & TICKETS
IMDbPro
IMDb Resume
Home | Top Movies | Photos | Independent Film | GameBase | Browse | Help
search All 300 go more | tips
Media from 300 (2006)
Photos (see all 108 | slideshow)
Videos (see all 14 videos)
Popular Titles (Displaying 1 Result)
1. 300 (2006)
aka "300: The IMAX Experience" - USA (IMAX version)

Web Page Templates

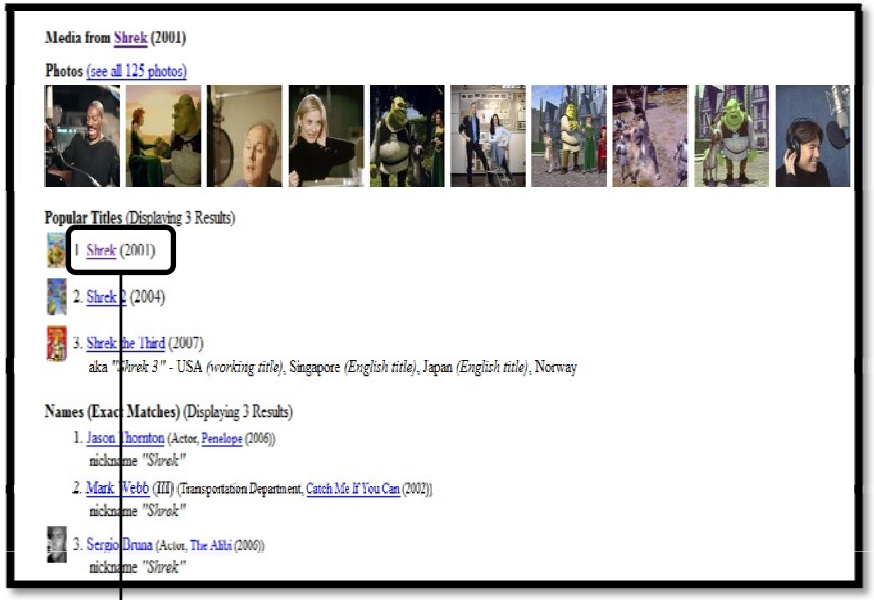


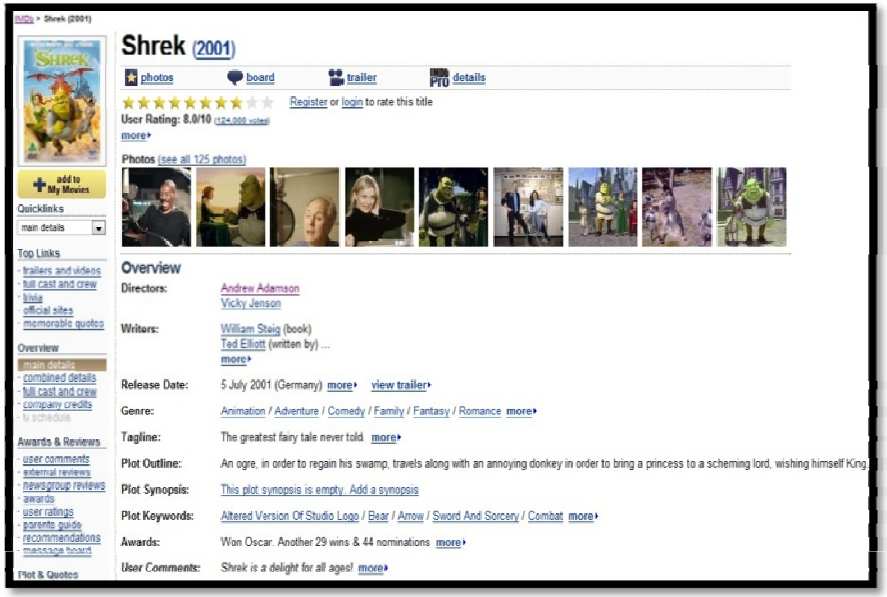
- Fix part (scaffolding) + dynamic part (data)
- Data is added dynamically from backend DB

We can identify constant HTML-Element (keyword), e.g. text (in DIV, H2, ...)


Images from Shrek (2001) (see all 63 photos)
Popular Titles (Displaying 3 Results)
1. Shrek (2001)
2. Shrek 2 (2004)
3. Shrek the Third (2007)
characteristic

To Result Page





execute actions

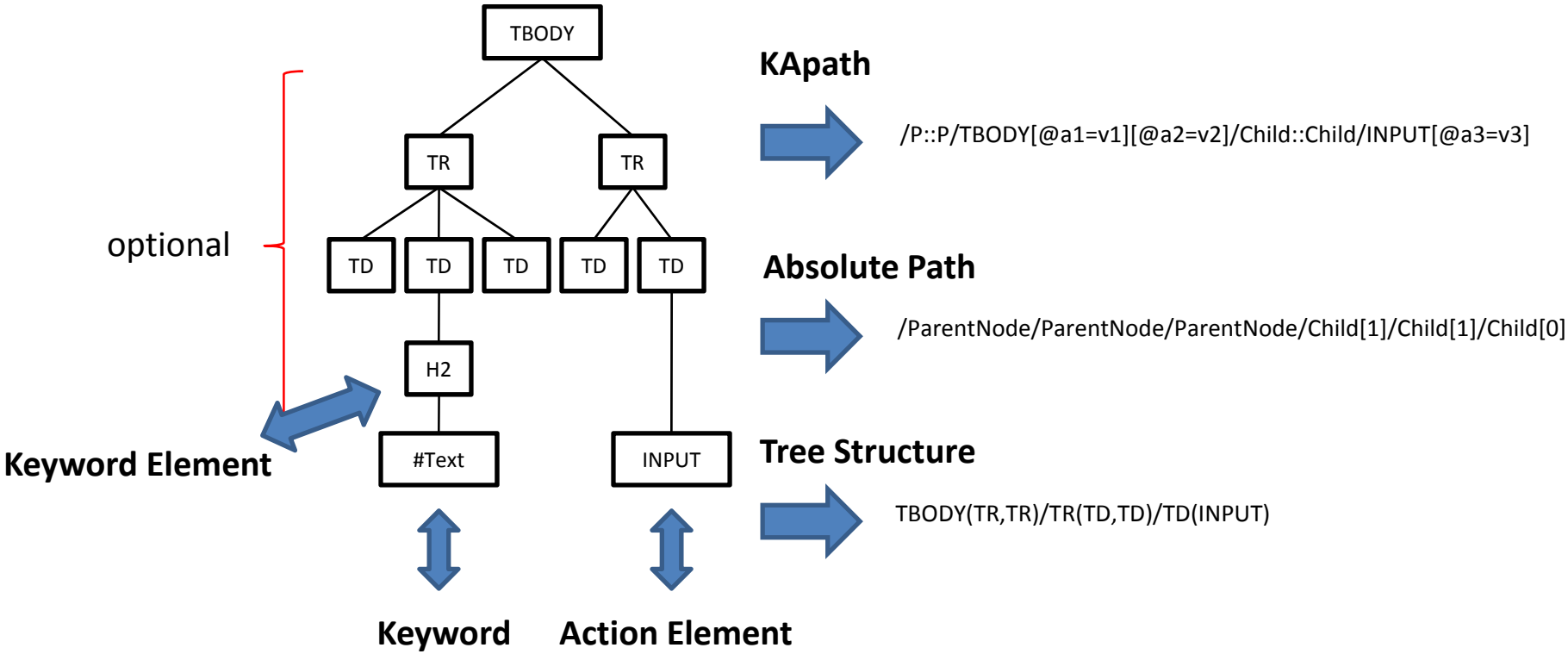
click link

Actions

- Keyword element
  - HTML-element that contains keyword-text
  - identify by click on the keyword-text
- Action element
  - HTML-element that relevant to performed action
  - monitored by system
- How can we address action element

  KApath
  - automatically generated
  - path from keyword element to action element
  - similar to Xpath
  - addtional: absolute path, tree structure

Addressing Action Elements



Evaluation

- 100 tested websites
  - Check of dynamic dependencies → 99% (from 0.5 to 30 seconds)
  - Deep Web Navigation → 96% (from 2.26 to 11.22 seconds)
- Open Issues
  - Delayed AJAX interactions
  - Session IDs
  - ...

Summary

- New Navigation Paradigm
  - Page-oriented
  - Short navigation expression
- Future Work
  - More resilient determination of intermediate pages