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
**Diagrams of Dynamic Dependencies**

1. **Diagram 1:**
   - **Marke:** BMW
   - **Modell:** Alle...
   - **Kraftstoff:** Alle...
   - **Angebote:** Händler- & Privatangebote
   - **Umkreis:** Alle... um PLZ
   - **Ganz Europa**

2. **Diagram 2:**
   - **Marke:** Mercedes-Benz
   - **Modell:** Alle...
   - **Kraftstoff:** Alle...
   - **Angebote:** Händler- & Privatangebote
   - **Umkreis:** Alle... um PLZ
   - **Ganz Europa**

**Dynamic Dependencies:**

- Marke and Modell are dynamic dependencies as they influence the available options for Kraftstoff and Angebote.
- Umkreis options change dynamically based on the selected Marke and Modell.

---

**Notes:**

- The diagrams illustrate the dynamic relationships between different categories and options in the context of a car marketplace scenario.
- The user interface is designed to filter search results based on the selected car model and type.
Dynamic or Static?

Idea: Simulate HTML-events

1. Save initial status of marked dropdown menus (temporarily)
   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

Attribute
....

Radius

5 km
10 km

Price-From

500 €
1000 €

Price-To

500 €
1000 €

Attribute name

Options

....

dynamic

Car-Brand

Audi

A1
A2

BMW

320
325

VW

Golf
Polo

car-Brand

car-Model
```
2 Deep Web Navigation
Movie-Name = "The Game Plan"

Movie-Name = "Shrek"

Intermediate-Page

Result-Page
Navigation Process

Form Field

label_1 → value_1
label_2 → value_2
label_3 → value_3
Search!

HTML page

Found Keyword?

yes → Execute Actions
no → Goal Site

Execute Actions

Return URL
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, ...)

Movie name: Shrek 2
Directors: Andrew Adamson, Vicky Jenson
Release Date: 5 July 2001
Images from Shrek (2001) (see all 63 photos)

Popular Titles (Displaying 3 Results)
1. Shrek (2001)

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

KPath
  • automatically generated
  • path from keyword element to action element
  • similar to Xpath
  • additional: absolute path, tree structure
Addressing Action Elements

Keyword Element

Optional

Tree Structure

Keyword

Action Element

Absolute Path

/K::P/TBODY[@a1=v1][@a2=v2]/Child::Child/INPUT[@a3=v3]

KPath

/TBODY(TR,TR)/TR(TD,TD)/TD(INPUT)
Evaluation

• 100 tested websites
  • Check of dynamic dependencies \( \rightarrow 99\% \) (from 0.5 to 30 seconds)
  • Deep Web Navigation \( \rightarrow 96\% \) (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