Wrapper Induction

Fabian M. Suchanek
Semantic IE

Source Selection and Preparation

Entity Recognition

Entity Disambiguation

Instance Extraction

Fact Extraction

Reasoning

You are here

singer Elvis
Generated Web pages

Web page generation is the process of producing several similar Web pages from a KB.
Example: Generated Web pages

- **Price**: 10 USD
- **Author**: William Steig

"Shrek"

by W. Steig

only 10 USD!

Buy it!
IE aims to reconstruct the KB

"Shrek"
by W. Steig
only 10 USD!
Buy it!
Def: Wrapper

A wrapper for a set of pages generated from the same KB is a function that extracts strings from such a page.

(Technically, it is the inverse function of the function that generated the page. The strings still have to be disambiguated and put in relation to yield facts.
Different applications have different more specific definitions of the "strings".)

Kushmerick: Wrapper Induction

"Shrek...", "90 min", "7.9"
Information is always in the same place

If we understand this...

then we understand this.

**Shrek - Der tollkühne Held**
(2001)
"Shrek" (original title)
Rated 7.9/10 from 294,865 users. Metascore: 84/100
Reviews: 9,189 user | 205 critic | 34 from Metacritic.com

An ogre, in order to regain his swamp, travels along with an annoying donkey in order to bring a princess to a scheming lord, wishing himself king.

Directors: Andrew Adamson, Vicky Jenson
Writers: William Steig (book), Ted Elliott (screenplay), 6 more credits
Stars: Mike Myers, Eddie Murphy, Cameron Diaz | See full cast and crew

---

**Elvis: Aloha from Hawaii**
(TV 1973)
TV Special - 87 min - Documentary | Music

Your rating: 1/10 from 866 users
Reviews: 30 user | 3 critic

A 1973 concert by Elvis Presley taped at the Convention Center in Honolulu, Hawaii. This was the first program to ever be beamed around the world by satellite.

Directors: Marty Pasetta, Gary Hovey, 1 more credit
Stars: Elvis Presley, James Burton, Jerry Scheff | See full cast and crew
Def: XPath

XPath is a formal language for selecting nodes in an XML document. [Wikipedia/XPath]

/ identifies the root node
K/T[i] identifies the i-th child with tag T of the node identified by K
K/T is K/T[1] if K has one T child

<html>
  <body>
    <h1>Aloha from Hawaii</h1>
    <p>This is a really great movie</p>
    <p>Stars:<i>Elvis Presley</i></p>
  </body>
</html>

Try it out

/html/body/p[2]/i

(XPath has many more expressions, allowing the union of node sets, the selection by attributes, etc.)
Task: XPath

Write XPath expressions that identify nodes whose text is ”Shrek”, ”W. Steig”, and ”84 min”.

<html>
  <body>
    <b>Shrek</b>
    <ul>
      <li>Creator: <b>W. Steig</b></li>
      <li>Duration: <i>84 min</i></li>
    </ul>
  </body>
</html>
Def: Wrapper induction

Wrapper induction is the process of generating a wrapper from a set of Web pages with strings to be extracted.

Web page

"Shrek", "7.9"

= 

/html/body/h1

/html/body/p[2]/i

Wrapper
Wrapper induction

Wrapper Induction requires as input Web pages with strings to be extracted. These can come, e.g.,
- from a KB

\[ \text{hasTitle(ShrekMovie, "Shrek")} \]

- from manual extraction

- from manual annotation in a GUI
Def: Wrapper Application

Wrapper application is the process of extracting its strings from a Web page.

Web page

+ Wrapper

= Strings

"Elvis", "11"
Wrapper application is the process of extracting its strings from a Web page.

- Wrapper application
- Web page
- Wrapper
- Strings
- Facts

Disambiguation, + relation

hasActor(e42, ElvisPresley)
hasRating(e42, "11.0")
Detail Pages & List Pages

Wrappers can be learned across several detail pages:

Wrappers can also be learned across items in a list:
Data may exhibit structure

Dronkeys:

<ul>
  <li>Eclair: female</li>
  <li>Bananas: flexible</li>
</ul>

Shrek’s kids:

<ul>
  <li>Farkle: male</li>
  <li>Fergus: male</li>
</ul>

family: tuple (
  name: string
  children: set (
    child: tuple (name: string, gender: string)))
ROADRUNKER: Learn types

ROADRUNKER is a system that can learn the Web page structure.

Page 1:
<ul>
<li>Peanut</li>
</ul>

Page 2:
<ul>
<li>Charles</li>
</ul>

Wrapper:
<ul>
<li>[FIELD]</li>
</ul>
ROADRUNNER: Learn types

ROADRUNNER is a system that can learn the Web page structure.

Page 1:
<ul>
<li>Peanut</li>
</ul>

Page 2:
<ul>
<li>Charles</li>
<li>Anne</li>
</ul>

Wrapper:
<ul>
<li>(<li>[FIELD])+</li>
</ul>
The Dronkeys<br>

Eclair: female<br>
Bananas: male<br>
Peanut: unknown<br>

For the WIEN system, a wrapper is of the form

<head, tail, left, right, left, right, ...>
<p>, <hr>, n, :, , <br>, ...

>end
Running WIEN

<html>
<body>
The Dronkeys<br>
<p>Eclair: female<br>Bananas: male<br>Peanut: unknown<br></p>
<hr>

For the WIEN system, a wrapper is of the form

<head, tail, left, right, left, right, ...>
<p>, <hr>, n, :, , <br>, ...>

1. Scroll to head
2. Scroll to left
3. Extract until right
4. Repeat until tail
Delimiters may differ

<html>
<body>
The Dronkeys<br/>
<p>
Eclair: female<br/>
Bananas (male)<br/>
Peanut is hybrid<br/>
<hr>

The STALKER system allows disjunctions in wrappers, such as:
The start of the gender is either “:” or “(”.

</p></body></html>
References

Kushmerick: Wrapper Induction
ROADRUNNER
Web data mining class
Muslea: STALKER