Interactive Web Data Extraction with WebLearn

Michal Ceresna, Max Goebel

February 8th, 2007

Agenda

- **Introduction, Objectives**
- Navigation
- Clustering
- Web Wrapping
  - Attribute Classifier
  - Compound Filter Learning
- Conclusion

Introduction

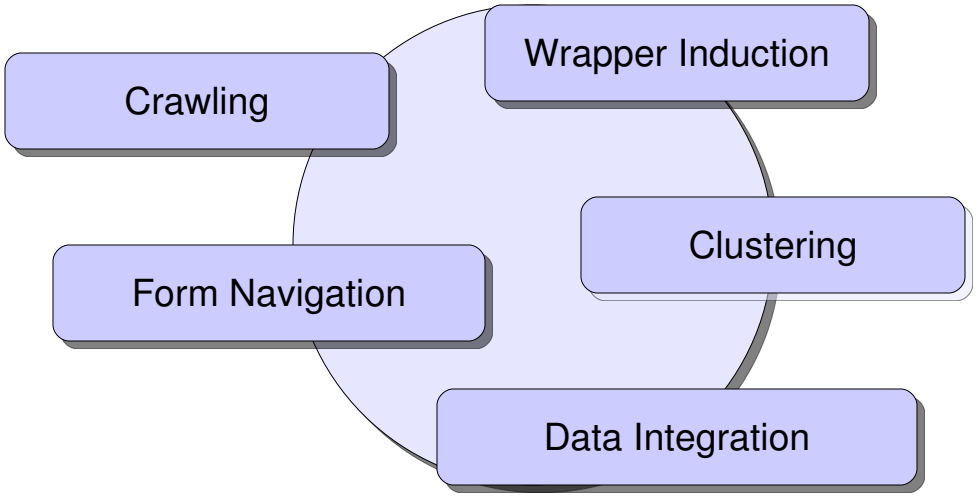
- Information Extraction from Web documents
- Data Mining Problems
  - Semi / unstructured HTML documents
  - Extraction patterns
  - Hidden Web
- Applications
  - market monitoring, price comparison, customer opinion, quality management

WebLearn Objectives

- Build integrated solution
- Interactive Wrapper Induction (Learning)
  - Combine WS and WI approaches
  - WI techniques to speed-up WS
  - WS techniques to speed-up WI
- Deep Web navigation
  - Interactive record & replay using JavaScript forms
- Data preprocessing (clustering)

WebLearn Workflow

- Wrapper Induction as one of many system modules
- Workflow model captures process workflows
- Driven by (yet encapsulated from) user



Agenda

- Introduction, Objectives
- **Navigation**
- Clustering
- Web Wrapping
  - Attribute Classifier
  - Compound Filter Learning
- Conclusion

Deepp Web Navigation

- Types
  - Crawling
  - Given navigation
  - Auto navigation
- Obstacles
  - Dynamic contents
  - Sessions, state-full
  - Authorizations, HTTPS
  - Proxies
  - AJAX
  - Page Interactions
    - Forms
    - In-page navigation
    - Flash

Record and Replay Navigations

- Synchronized workflow model
  - forking, conditional branching, etc.
- Scriptable actions (intuitive primitives)
  - find, go, ...
- Recorded on-the-fly from user interaction

Agenda

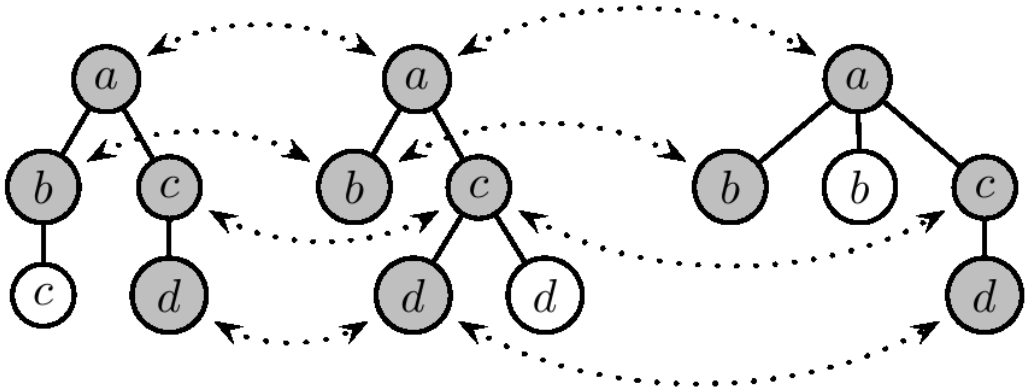
- Introduction, Objectives
- Navigation
- **Clustering**
- Web Wrapping
  - Attribute Classifier
  - Compound Filter Learning
- Conclusion

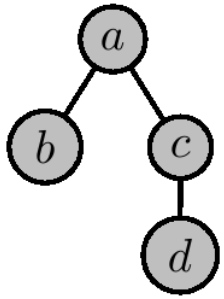
Clustering

- Heterogeneous input data from IR
- Group into *similarity clusters*
- Similarity of DOM tree structures

Tree Edit Distance

- Efficient tree matching algorithm
- Gives score on structural similarity of two trees





Template Tree

Agenda

- Introduction, Objectives
- Navigation
- Clustering
- **Web Wrapping**
  - Attribute Classifier
  - Compound Filter Learning
- Conclusion

Terminology

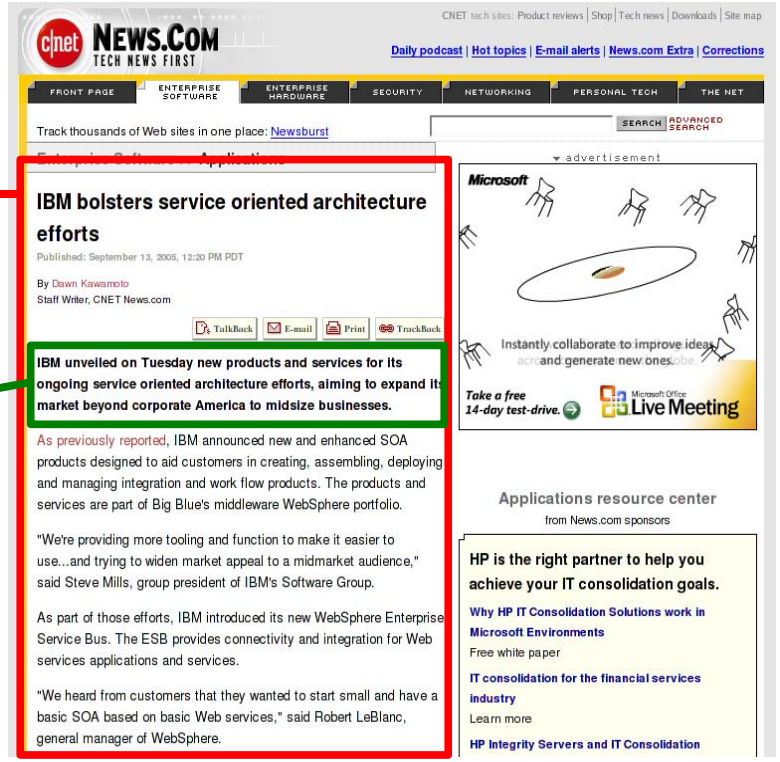
Wrapper	Hierarchically organized structure of patterns
Pattern	Container for pieces of information with the same meaning
Filters	Rules defining how to extract information into patterns

Nested Pattern Hierarchies

Patterns

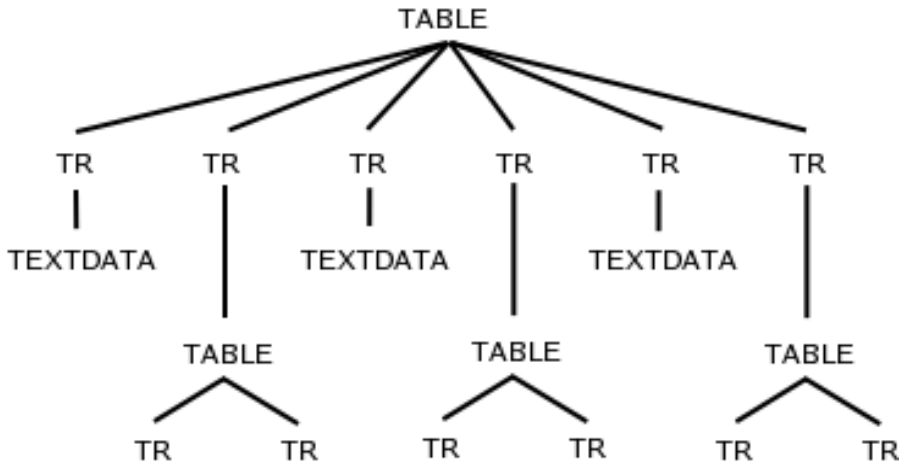
Article

- Title
- Abstract
- Content
  - Paragraph



Filters

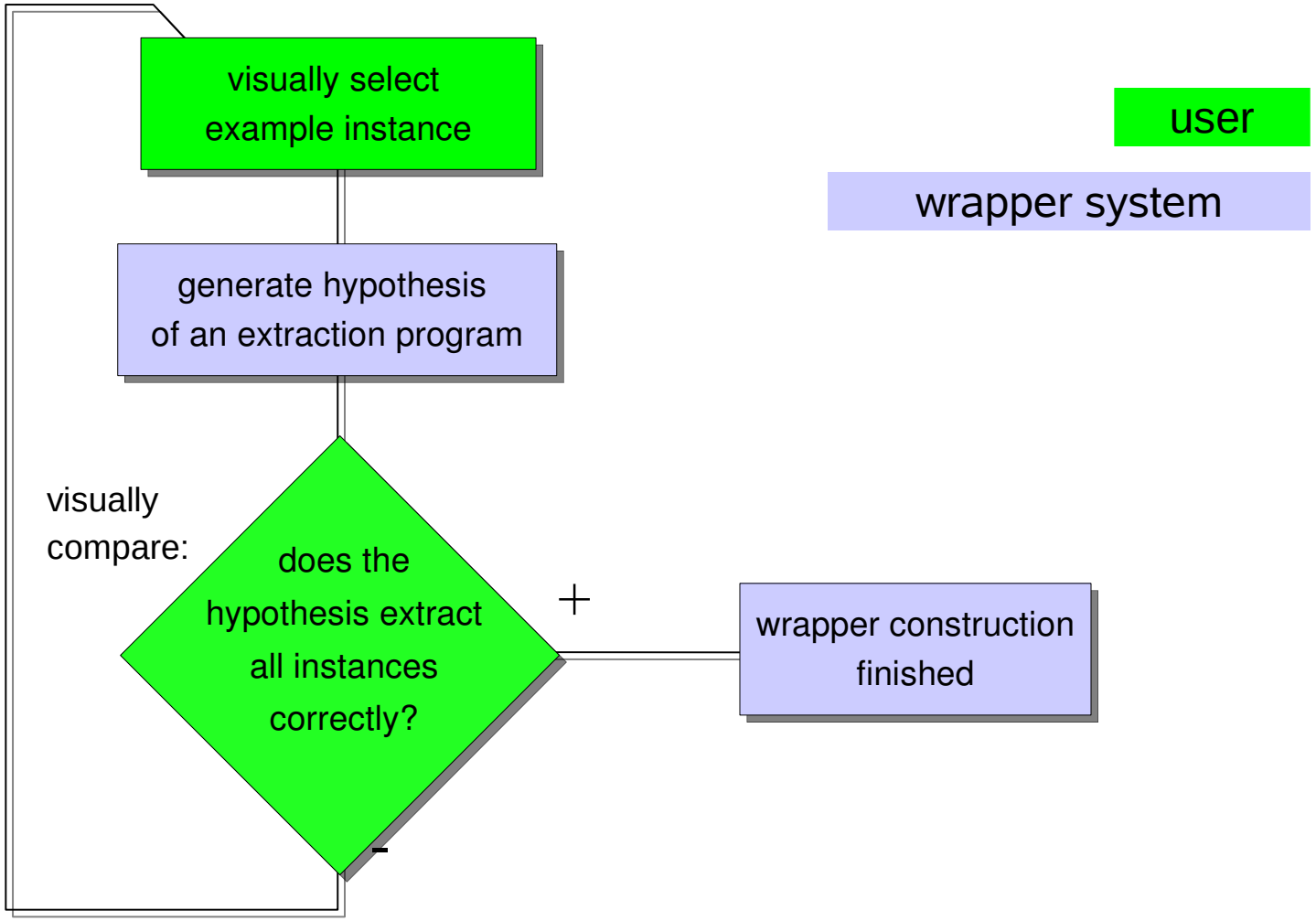
- Extraction Rules
- Node selection function for DOM trees
- Example:



Web Wrappers

- Wrapper Specification (WS)
  - Manual labelling training data
  - Cumbersome (domain/programming prerequisites)
  - High level of control
- Wrapper Induction (WI)
  - Semi-supervised
  - Fully-automatic approaches
  - Application domains vary

Interactive Learning



Agenda

- Introduction, Objectives
- Navigation
- Clustering
- Web Wrapping
  - **Attribute Classifier**
  - Compound Filter Learning
- Conclusion

Attribute Classification

- Learning from tree structure and **attributes** in the DOM tree
- Allows to use:
  - built-in knowledge of HTML standard
  - features known from existing wrapper systems such as syntactic and semantic concepts

date, number, city, country

Attribute Features

- Iterating over instances in clusters compute features based on attribute values

BL3_href_val	BL3_href_protocol	BL3_bgcolor_val	BL4_text_val	extract*
http://...	http	black	contact1	yes
mailto:...	mailto	!missing	contact2	no
http://...	http	black	contact3	yes

BL5_text_val	BL5_text_isCity	extract*
Prague	yes	yes
Austria	no	no
Vienna	yes	yes

Attribute Learning Algorithm

- Decision tree is constructed for feature table
- XPath expression finds all potential instances with correct tree shape for the current cluster
- Decision tree classifier filters instances with correct attributes

Cluster1 ID3:
BL3_bg_color != '!missing' : yes

Cluster2 ID3 :
BL5_text_isCity = 'yes' : yes

Results

Source	URL	Pattern	Examples
Amazon Camera List	http://www.amazon.com/exec/obidos/tg/browse/...	Camera	1+2
Google Search	http://www.google.at/search?...	SearchResult	2+0
Yahoo Email Search	http://email.people.yahoo.com/py/psEmailSearch.py?...	PeopleEntry	1+0
IMDb Title Details	http://imdb.com/title/...	Actor	1+0
IMDb Title Details	http://imdb.com/title/...	Director	1+3
Excite Weather	http://my.excite.com/weather/obs.jsp?...	Forecast	2+1

Agenda

- Introduction, Objectives
- Navigation
- Clustering
- Web Wrapping
  - Attribute Classifier
  - **Compound Filter Learning**
- Conclusion

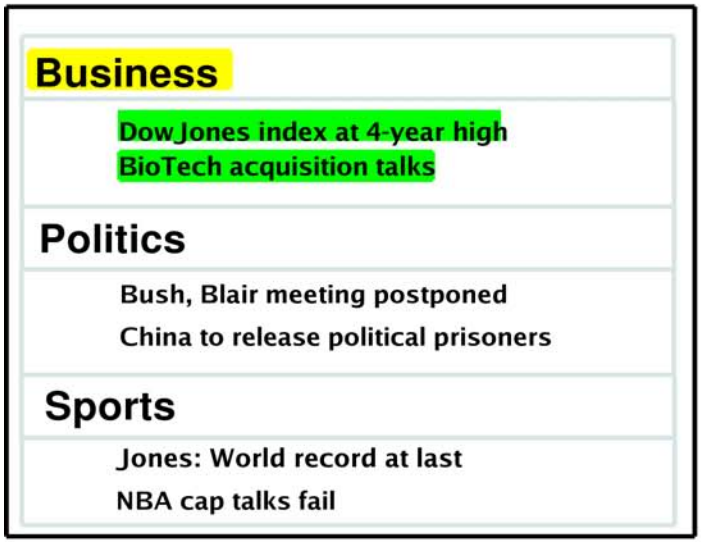
Compound Filter Learning

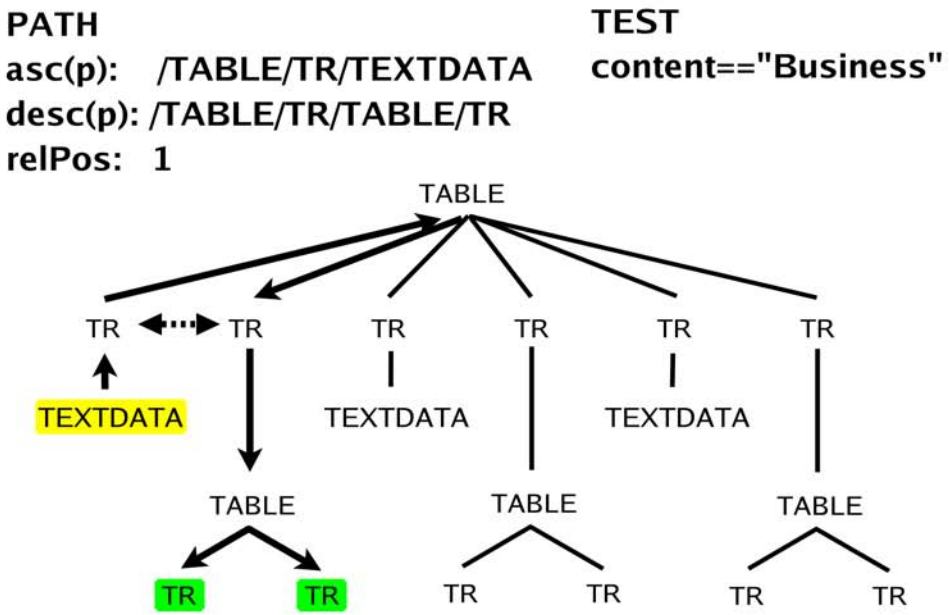
- Filters based on remote reference objects
- Specify relation from one node in document to another
- Simple, but expressive semi-automatic approach
- Well-suited for interactive IE systems
- Formalism of filters similar to human concepts
- Wrappers are easy to understand and interpret

Filter Definition

- Filters defined by Paths and Tests

$$C(n) \Leftrightarrow \exists n' : path(C)(n, n') \wedge test(C)(n')$$

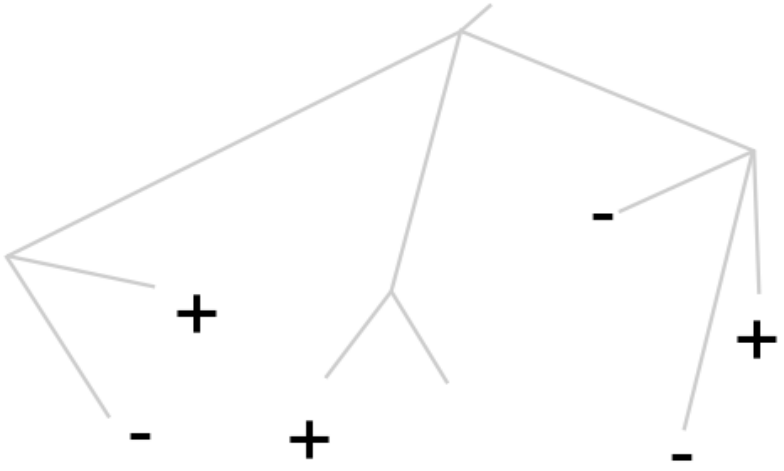




Compound Filters

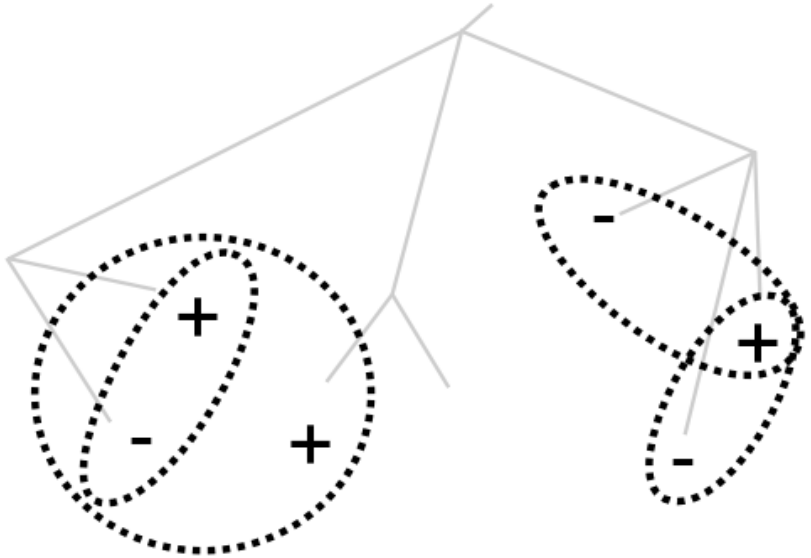
- Combination of filters using operators: $\wedge, \vee, \neg$.
- Case of k – DNF learning:
  - generate all conjunctions of literals with size < k
  - drop all conjunctions described by negative examples
  - get “optimal conjunction” from minimal covering algorithm

Filter Learning Algorithm



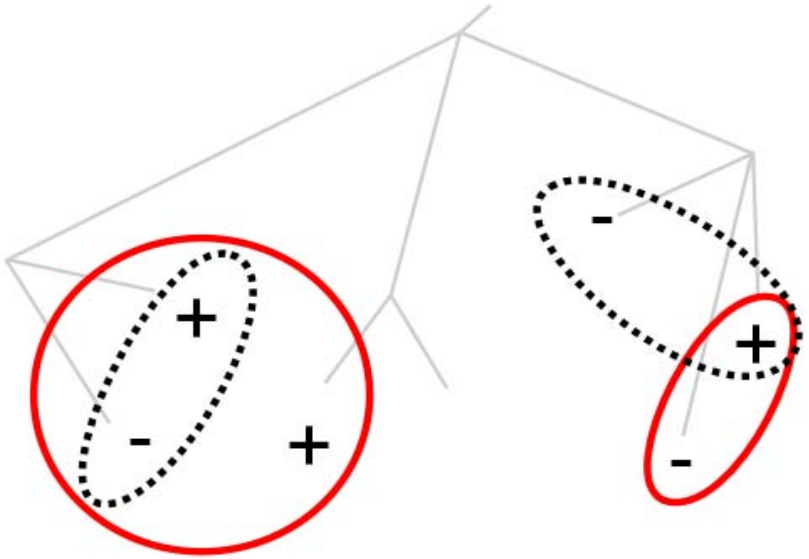
- All filters consister
- Selection of minimum descriptive filter set
  - Filter dominance, filter equivalence
- Combining filters to concept using k-DNF algorithm

Filter Learning Algorithm



- All filters consister
- Selection of minimum descriptive filter set
  - Filter dominance, filter equivalence
- Combining filters to concept using k-DNF algorithm

Filter Learning Algorithm



- All filters consister
- Selection of minimum descriptive filter set
  - Filter dominance, filter equivalence
- Combining filters to concept using k-DNF algorithm

Results

Corpus ID	Description	# total docs	# docs used	# interactions
Okra	Name/Address search	250	1.33	2.33
Bigbook	Yellow Pages	235	1.0	2.0
Yahoo	Directory search	84	3.81	8.36
NYTimes	Newspaper articles	10	1.0	1.0
Google	Search engine results	33	1.0	2.0
Slashdot	News for nerds	19	1.8	2.5
CiteSeer	Reference catalogue	41	5.0	9.33
Ebay	Shopping portal	34	1.81	3.2
LeMonde	Newspaper articles	43	1.0	1.0

Agenda

- Introduction, Objectives
- Navigation
- Clustering
- Web Wrapping
  - Attribute Classifier
  - Compound Filter Learning
- **Conclusion**

Conclusion

- Integrated system combining
  - Navigation
  - Clustering
  - Extraction aspects
- Different wrapper learning approaches
  - Attribute versus structure-based
- Current Work
  - Query Learning