Towards the Discovery of Implicit Metadata in Commercial Web Pages

Vojtěch Svátek, Martin Kavalec,
Jiří Klemperer

University of Economics, Prague
Department of Information and Knowledge Engineering

Implicit Metadata

- *Implicit metadata* are not separate from the document content but have to be *extracted* from it by means of heuristics

- Static websites of small companies seem to be hot candidates for such extraction

- We attempt to distinguish between extraction from *free-text sentences* (exploiting the parse-tree) and from *text broken down to structural HTML elements* (exploiting the mark-up tree)

- Both types of structures could be mapped on basic *RDF triples* !

Free-text Sentences

- Typically description of the *company profile* incl. *generic products/services*

- **Heuristic:** target *metadata* (themselves domain-dependent) are often linked as syntactical object to a *domain-neutral* term or phrase – the linguistic *indicator* !

- Examples of simple indicators: *supply*, *offer*, *specialize in*, *sell*, *manufacture*, *buy*, *provide*, *find*, *need*...

- Indicators vary in 'accuracy' i.e. probability to be linked to desirable metadata

Acquisition and Evaluation of Linguistic Indicators

- **Hypothesis:** target metadata on the page often coincide with terms from *web directory headings*

  `.../Manufacturing/Materials/Metals/Steel/...`

- By comparing the headings with the content of pages referenced by the given directory node, we can *automatically* tag the metadata in these pages

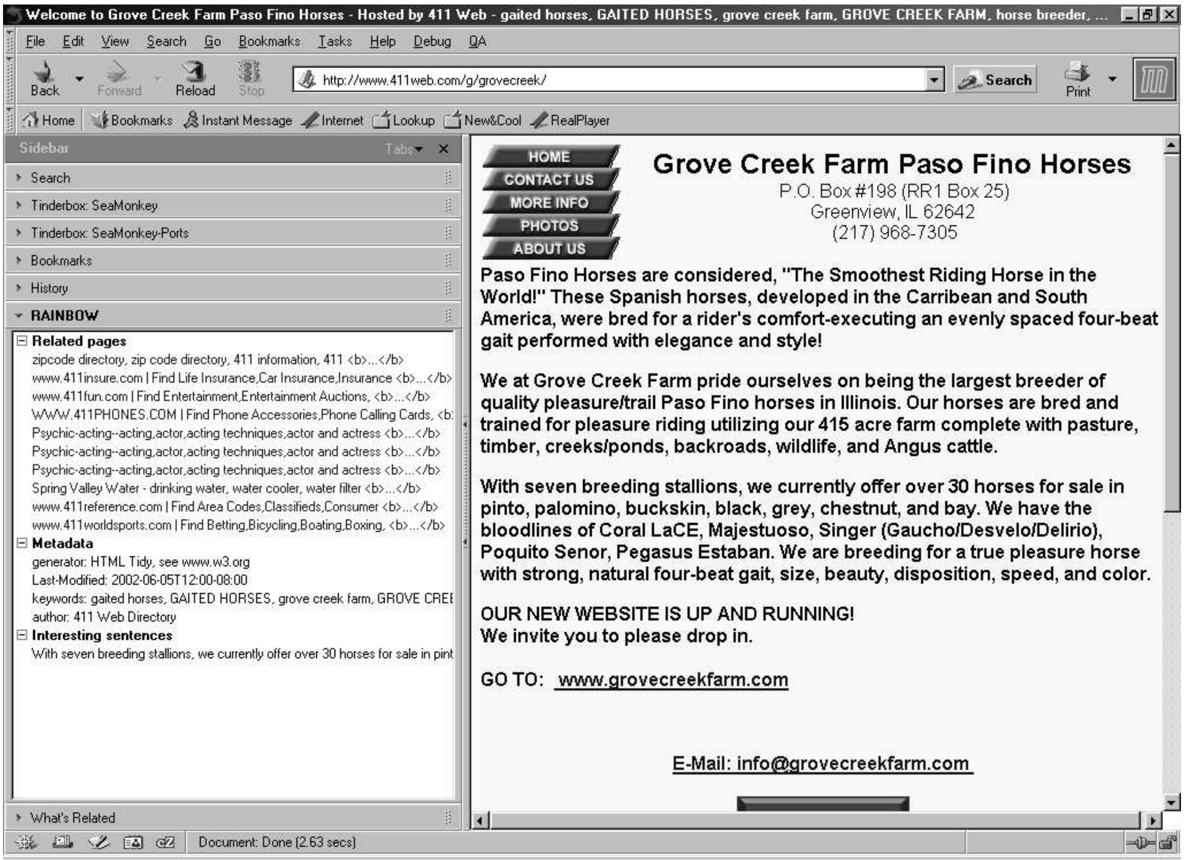
- Parse-trees of sentences with tagged metadata become *training cases* for learning the indicators

- Evaluation of accuracy and completeness of indicators requires user assessment; some results:

indicator	–	?	+	precision
include	8	4	18	60–73%
provide	9	3	28	70–78%
offer	6	1	21	75–79%
specialize	0	1	18	95–100%
(other)	3	5	5	38–77%
total	26	14	90	77–80%

Example Mapping to RDF

(not yet implemented...)



RDF subject: `http://www.411.web.com/g/grovecreek/`

RDF predicate: could be ontology concept ‘indication-of-generic-product’ linked to a collection of lexical indicators including ‘offer’

RDF object: ‘horses’ or ‘horses in pinto, palomino...’

(further linguistic heuristics needed for extraction !)

HTML Mark-up Structures – General View

- Several types of commercial information typically embedded in structural mark-up: addresses, prices, product codes...

- Other occur alternatively in free text or in mark-up: generic products, lists of projects or business partners...

- **But**: the use of HTML elements conditioned by the *outlook in browser* rather than by the semantics intended by W3C...!

- **However:** the number of ways how to obtain the desired outlook given a *fixed set of HTML constructs* is not unlimited !

HTML Mark-up Structures – Hypothesis

- Relevant micro-level structures of HTML (plus some special symbols) correspond to *property* - *value* pairs

- The *property* part often contains a *domain-neutral indicator* (usually a noun phrase) while the *value* part contains the *target metadata* !

- Simple examples of micro-level structures:
  - heading + paragraph
  - pairs of cells in a two-column table
  - font-emphasized text ending with colon + subsequent plain text

Future Work

- Implementation of *actual metadata extraction heuristics* for *linguistic indicators*

- Systematic evaluation of the hypothesis on *micro-level HTML-structures* and subsequent implementation of heuristics

- Exploitation of the resulting knowledge bases in the *Rainbow* architecture for multiway metadata acquisition, see

  `http://rainbow.vse.cz`

- Elaboration on *semantic-web* issues:

  - determination of RDF subjects

  - reuse of RDF predicates
    (linked to lexical indicators)