
Information Extraction

Information Extraction (IE)

- Identify specific pieces of information (data) in a unstructured or semi-structured textual document.
- Transform unstructured information in a corpus of documents or web pages into a structured database.
- Applied to different types of text:
 - Newspaper articles
 - Web pages
 - Scientific articles
 - Newsgroup messages
 - Classified ads
 - Medical notes

MUC

- DARPA funded significant efforts in IE in the early to mid 1990's.
- Message Understanding Conference (MUC) was an annual event/competition where results were presented.
- Focused on extracting information from news articles:
 - Terrorist events
 - Industrial joint ventures
 - Company management changes
- Information extraction of particular interest to the intelligence community (CIA, NSA).

Other Applications

- Job postings
- Job resumes
- Seminar announcements
- Company information from the web
- Apartment rental ads
- Molecular biology information from MEDLINE

Sample Job Posting

Subject: **US-TN-SOFTWARE PROGRAMMER**
Date: **17 Nov 1996** 17:37:29 GMT
Organization: Reference.Com Posting Service
Message-ID: <**56nigp\$mrs@bilbo.reference.com**>

SOFTWARE PROGRAMMER

Position available for Software Programmer experienced in generating software for PC-Based **Voice Mail** systems. Experienced in **C** Programming. Must be familiar with communicating with and controlling voice cards; preferable Dialogic, however, experience with others such as Rhetorix and Natural Microsystems is okay. Prefer **5** years or more experience with **PC Based Voice Mail**, but will consider as little as **2** years. Need to find a Senior level person who can come on board and pick up code with very little training. Present Operating System is **DOS**. May go to **OS-2** or **UNIX** in future.

Please reply to:
Kim Anderson
AdNET
(901) 458-2888 fax
kimander@memphisonline.com

Extracted Job Template

computer_science_job
id: 56nigp\$mrs@bilbo.reference.com
title: SOFTWARE PROGRAMMER
salary:
company:
recruiter:
state: TN
city:
country: US
language: C
platform: PC \ DOS \ OS-2 \ UNIX
application:
area: Voice Mail
req_years_experience: 2
desired_years_experience: 5
req_degree:
desired_degree:
post_date: 17 Nov 1996

Amazon Book Description

```
....  
</td></tr>  
</table>  
<b class="sans">The Age of Spiritual Machines : When Computers Exceed Human Intelligence</b><br>  
<font face=verdana,arial,helvetica size=-1>  
by <a href="/exec/obidos/search-handle-url/index=books&field-author=  
Kurzweil%2C%20Ray/002-6235079-4593641">  
Ray Kurzweil</a><br>  
</font>  
<br>  
<a href="http://images.amazon.com/images/P/0140282025.01.LZZZZZZZ.jpg">  
</a>  
<font face=verdana,arial,helvetica size=-1>  
<span class="small">  
<span class="small">  
<b>List Price:</b> <span class=listprice>$14.95</span><br>  
<b>Our Price: <font color=#990000>$11.96</font></b><br>  
<b>You Save:</b> <font color=#990000><b>$2.99 </b>  
(20%)</font><br>  
</span>  
<p> <br>...
```

Extracted Book Template

Title: **The Age of Spiritual Machines :**
When Computers Exceed Human Intelligence

Author: **Ray Kurzweil**

List-Price: **\$14.95**

Price: **\$11.96**

:
:

Web Extraction

- Many web pages are generated automatically from an underlying database.
- Therefore, the HTML structure of pages is fairly specific and regular (*semi-structured*).
- However, output is intended for human consumption, not machine interpretation.
- An IE system for such generated pages allows the web site to be viewed as a structured database.
- An extractor for a semi-structured web site is sometimes referred to as a *wrapper*.
- Process of extracting from such pages is sometimes referred to as *screen scraping*.

Template Types

- Slots in template typically filled by a substring from the document.
- Some slots may have a fixed set of pre-specified possible fillers that may not occur in the text itself.
 - Terrorist act: threatened, attempted, accomplished.
 - Job type: clerical, service, custodial, etc.
 - Company type: SEC code
- Some slots may allow multiple fillers.
 - Programming language
- Some domains may allow multiple extracted templates per document.
 - Multiple apartment listings in one ad

Simple Extraction Patterns

- Specify an item to extract for a slot using a regular expression pattern.
 - Price pattern: “\b\$\d+(\.\d{2})?\b”
- May require preceding (pre-filler) pattern to identify proper context.
 - Amazon list price:
 - Pre-filler pattern: “List Price: ”
 - Filler pattern: “\$\d+(\.\d{2})?\b”
- May require succeeding (post-filler) pattern to identify the end of the filler.
 - Amazon list price:
 - Pre-filler pattern: “List Price: ”
 - Filler pattern: “.+”
 - Post-filler pattern: “”

Pre-Specified Filler Extraction

- If a slot has a fixed set of pre-specified possible fillers, text categorization can be used to fill the slot.
 - Job category
 - Company type
- Treat each of the possible values of the slot as a category, and classify the entire document to determine the correct filler.

Natural Language Processing

- If extracting from automatically generated web pages, simple regex patterns usually work.
- If extracting from more natural, unstructured, human-written text, some NLP may help.
 - Part-of-speech (POS) tagging
 - Mark each word as a noun, verb, preposition, etc.
 - Syntactic parsing
 - Identify phrases: NP, VP, PP
 - Semantic word categories (e.g. from WordNet)
 - KILL: kill, murder, assassinate, strangle, suffocate
- Extraction patterns can use POS or phrase tags.
 - Crime victim:
 - Prefiller: [POS: V, Hypernym: KILL]
 - Filler: [Phrase: NP]

Learning for IE

- Writing accurate patterns for each slot for each domain (e.g. each web site) requires laborious software engineering.
- Alternative is to use machine learning:
 - Build a training set of documents paired with human-produced filled extraction templates.
 - Learn extraction patterns for each slot using an appropriate machine learning algorithm.
- Rapier system learns three regex-style patterns for each slot:
 - Pre-filler pattern
 - Filler pattern
 - Post-filler pattern

Evaluating IE Accuracy

- Always evaluate performance on independent, manually-annotated test data not used during system development.
- Measure for each test document:
 - Total number of correct extractions in the solution template: N
 - Total number of slot/value pairs extracted by the system: E
 - Number of extracted slot/value pairs that are correct (i.e. in the solution template): C
- Compute average value of metrics adapted from IR:
 - Recall = C/N
 - Precision = C/E
 - F-Measure = Harmonic mean of recall and precision

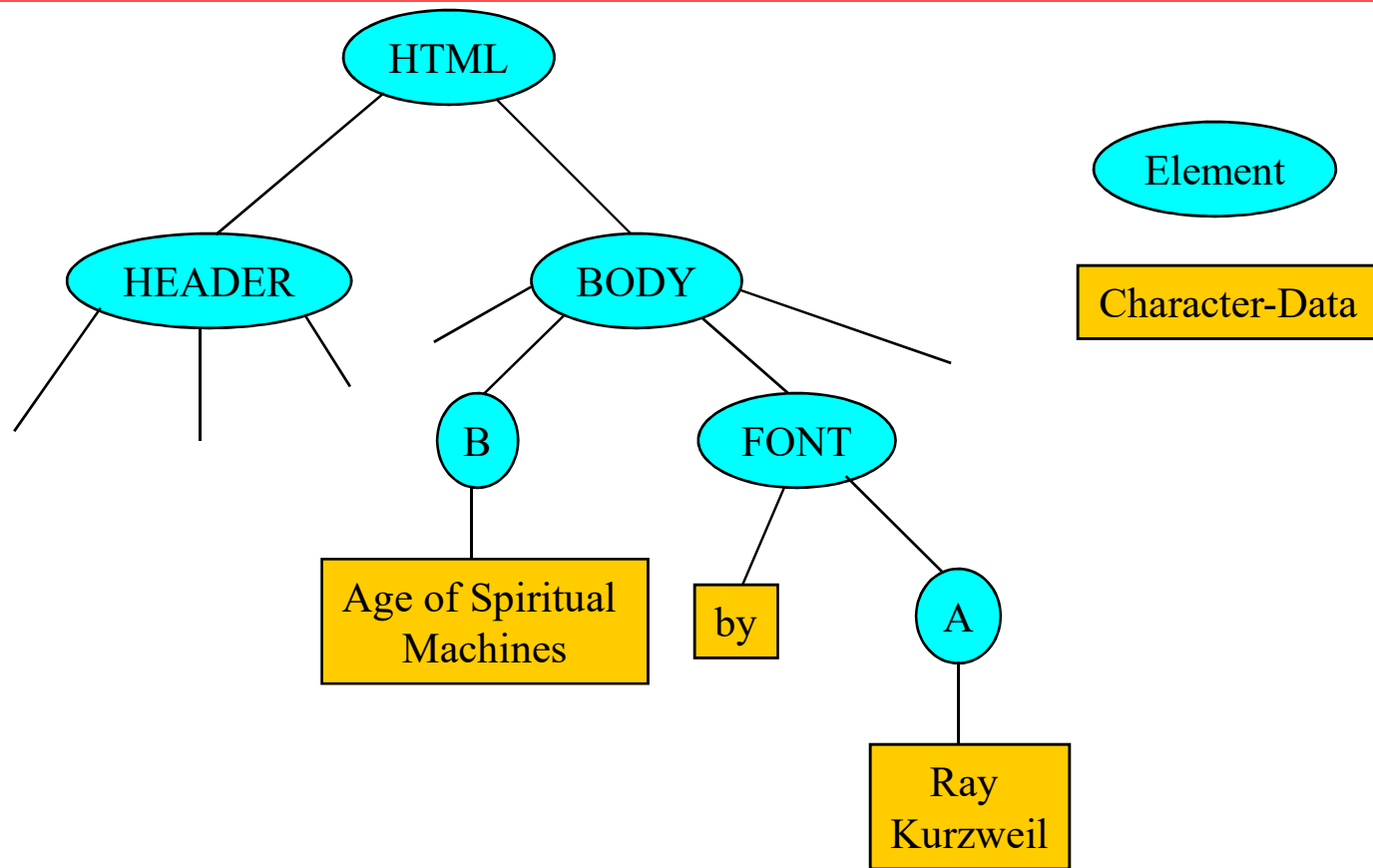
XML and IE

- If relevant documents were all available in standardized XML format, IE would be unnecessary.
- But...
 - Difficult to develop a universally adopted DTD format for the relevant domain.
 - Difficult to manually annotate documents with appropriate XML tags.
 - Commercial industry may be reluctant to provide data in easily accessible XML format.
- IE provides a way of automatically transforming semi-structured or unstructured data into an XML compatible format.

Web Extraction using DOM Trees

- Web extraction may be aided by first parsing web pages into DOM trees.
- Extraction patterns can then be specified as paths from the root of the DOM tree to the node containing the text to extract.
- May still need regex patterns to identify proper portion of the final CharacterData node.

Sample DOM Tree Extraction



Title: HTML → BODY → B → CharacterData

Author: HTML → BODY → FONT → A → CharacterData

Shop Bots

- One application of web extraction is automated comparison shopping systems.
- System must be able to extract information on items (product specs and prices) from multiple web stores.
- User queries a single site, which integrates information extracted from multiple web stores and presents overall results to user in a uniform format, e.g. ordered by price.

Information Integration

- Answering certain questions using the web requires integrating information from multiple web sites.
- Information integration concerns methods for automating this integration.
- Requires wrappers to accurately extract specific information from web pages from specific sites.
- Treat each wrapped site as a database table and answer complex queries using a database query language (e.g. SQL).

Information Integration Example

- Question: What is the closest theater to my home where I can see both “Doctor Strange” and “Jason Bourne”?
 - From movie listing site, extract theaters and their addresses where. are playing.
 - Intersect the two to find the theaters playing both.
 - Query map site for driving directions from your home address to the address of each of these theaters.
 - Extract distance and driving instructions for each.
 - Sort results by driving distance.
 - Present driving instructions for closest theater.