Informations-Extraktion

Information-Extraktion (IE)

- Identifiziere spezifische Informationen (Daten) in einem unstrukturierten oder halb-strukturierten Textdokument.
- Transformiere unstrukturierte Informationen eines Dokumentenkorpus oder von Webseiten in eine strukturierte Datenbank.
- Kann auf die folgenden Textarten angewendet werden:
  - Zeitungsartikel
  - Webseiten
  - Wissenschaftliche Artikel
  - Newsgroup-Beiträge
  - Zeitungsannoncen
  - Medizinische Beiträge

MUC

- DARPA förderte Anfang bis Mitte 1990 bedeutende Bemühungen im Bereich des IE.
- Die Message Understanding Conference (MUC) war eine jährliche Konferenz mit Wettbewerb, auf der Ergebnisse präsentiert wurden.
- Fokussiert auf die Extraktion von Informationen aus Nachrichtenartikeln:
  - Terroristische Ereignisse
  - Industrielle Joint Ventures
  - Änderungen im Firmen-Management
- Informations-Extraktion von besonderem Interesse für Geheimdienste (CIA, NSA).

weitere Anwendungen

- Job Postings:
  - Newsgroups: Rapier von austin.jobs
  - Webseiten: Flipdog
- Job Zusammenfassungen:
  - BurningGlass
  - Mohomine
- Seminar-Ankündigungen
- Firmeninformationen aus dem Web
- Universitäts-Informationen aus dem Web
- Anzeigen für Wohungsvermietungen
- Information zur Molekularbiologie von MEDLINE

Beispiel 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 Buch-Beschreibung

```
….
</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">
<img src="http://images.amazon.com/images/P/0140282025.01.MZZZZZZZ.gif" width=90
   height=140 align=left border=0></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>…
```

Extrahierte Buch Informationen

Title: The Age of Spiritual Machines :
When Computers Exceed Human Intelligence
Author: Ray Kurzweil
List-Price: $14.95
Price: $11.96
:
:

Web-Extraktion

- Viele Webseiten werden automatisch von einer zu Grunde liegenden Datenbank erzeugt.
- Daher ist die HTML-Struktur von Seiten ziemlich spezifisch und regelmäßig (*halb-strukturiert*).
- Jedoch ist der Output nur für den Menschen bestimmt und nicht für die Nutzung durch Maschinen.
- Ein IE System zeigt eine solche Webseite wie eine strukturierte Datenbank.
- Ein Extraktor für eine halb-strukturierte Webseite wird häufig als *Wrapper* bezeichnet.
- Der Extraktions-Prozess solcher Seiten wird als *Screen Scraping* bezeichnet.

Vorlagenarten (Template Types)

- Slots einer Vorlage werden typischerweise durch einen Substring aus dem Dokument gefüllt.
- Einige Slots können eine festgelegte Menge von vorspezifizierten möglichen Füllern haben, die möglicherweise nicht im Text selbst erscheinen.
  - Terroristischer Akt: angedroht, versucht, ausgeführt.
  - Jobart: Bürotätigkeit, Dienstleistung, Pflege, etc.
- Einige Slots können eine Vielzahl von Füller erlauben.
  - Programmiersprache
- Für einige Domainen benötigt man mehrere Vorlagen pro Dokument um die extrahierten Daten ablegen zu können.
  - Mehrere Apartments in einer Anzeige

Einfache Extraktions-Muster

- Spezifiziere ein zu extrahierendes Objekt für einen Slot unter Verwendung eines regulären Audrucks.
  - Price pattern: “\b\$\d+(\.\d{2})?\b”
- Kann vorangehende Muster (pre-filler) benötigen, um den passenden Kontext zu identifizieren.
  - Amazon Listenpreis:
    - pre-filler Muster: “List Price: ”
    - Füller Muster: “\$\d+(\.\d{2})?\b”
- Kann nachfolgende Muster (post-filler) benötigen, um das Ende des Füllers zu identifizieren.
  - Amazon Listenpreis:
    - pre-filler Muster : “List Price: ”
    - Füller Muster : “.+”
    - post-filler Muster: “”

Einfache Vorlagen-Extraktion

- Extrahiere Slots der Reihenfolge nach: beginn mit der Suche nach dem Füller des *n*+1. Slots an der Stelle, an der die Suche nach dem Füller für den Slot *n* endete. Geht immer von Slots in einer festgelegten Reihenfolge aus.
  - Titel
  - Autor
  - Listenpreis
  - …
- Muster sind spezifisch genug, um jeden Füller immer identifizieren zu können, wobei immer am Anfang des Dokuments begonnen wird.

Vorspezifizierte Füller-Extraktion

- Falls ein Slot eine festgelegte Menge von vorspezifizierten möglichen Füllern hat, kann Textkategorisierung verwendet werden, um den Slot zu füllen.
  - Jobkategorie
  - Firmentyp
- Behandle jeden der möglichen Werte des Slots als Kategorie und klassifiziere das ganze Dokument, um den korrekten Füller zu bestimmen.

Natürliche Sprachverarbeitung

- Einfache Regex-Muster funktionieren üblicherweise gut auf automatisch generierten Webseiten.
- Versucht man Daten von Dokumenten mit natürlichsprachlichem, unstrukturierten und menschlich-geschriebenem Text zu extrahieren, können NLP-Ansätze helfen.
  - Part-of-speech (POS) tagging
  - Syntactic parsing
    - Markiere jedes Wort als ein Nomen, Verb, Präposition, etc.
    - Identifiziere Phrasen: NP, VP, PP
  - Semantische Wortkategorien (z.B. von WordNet)
    - TÖTE: töte, Mörder, ermorden, erwürge, ersticke
- Extraktionsmuster können dann POS oder Phrasentags verwenden.
  - Kriminalopfer:
    - Vorfüller: [POS: V, Hypernym: KILL]
    - Füller: [Phrase: NP]

Lernen für IE

- Das Schreiben von exakten Mustern für jeden Slot für jede Domäne (z.B. jede Webseite) erfordert aufwändiges Software Engineering.
- Eine Alternative ist der Einsatz von maschinellem Lernen :
  - Bilde eine Trainingsmenge von Dokumenten, zusammen mit manuell erstellten, gefüllten Extraktions-Vorlagen.
  - Lerne Extraktionsmuster für jeden Slot unter Verwendung eines geeigneten Algorithmus' aus dem maschinellen Lernen.
- Rapier lernt drei Regex-Muster für jeden Slot:
  - Pre-filler Muster
  - Füllermuster
  - Post-filler Muster

Evaluierung der Genauigkeit bei der IE

- Evaluiere immer die Leistung an unabhängigen, manuell gekennzeichneten Testdaten, die nicht während der Systementwicklung verwendet wurden.
- Messe für jedes Testdokument:
  - Die Gesamtzahl der korrekten Extraktionen in der Lösungsvorlage : *N*
  - Die Gesamtzahl der Slot/Wertpaare, die vom System extrahiert werden: *E*
  - Anzahl der extrahierten Slot/Wertpaare, die korrekt sind (d.h. in der Lösungsvorlage): *C*
- Berechne den Durchschnittswert der aus dem IR angepassten Maße:
  - Recall = *C*/*N*
  - Precision = *C*/*E*
  - F-Measure = Harmonisches Mittel von Recall und Precision

XML und IE

- Falls relevante Dokumente im Standard-Format XML verfügbar wären, wäre IE unnötig.
- Aber…
  - Schwierig, ein universell angepasstes DTD-Format für die jeweilige Domäne zu entwickeln.
  - Schwierig, Dokumente mit geeigneten XML-Tags manuell zu annotieren.
  - Industrie ist zögerlich, Daten im leicht zugänglichen XML-Format zu liefern.
- IE liefert einen Weg des automatischen Transformierens von halb-strukturierten oder unstrukturierten Daten in ein XML-kompatibles Format.

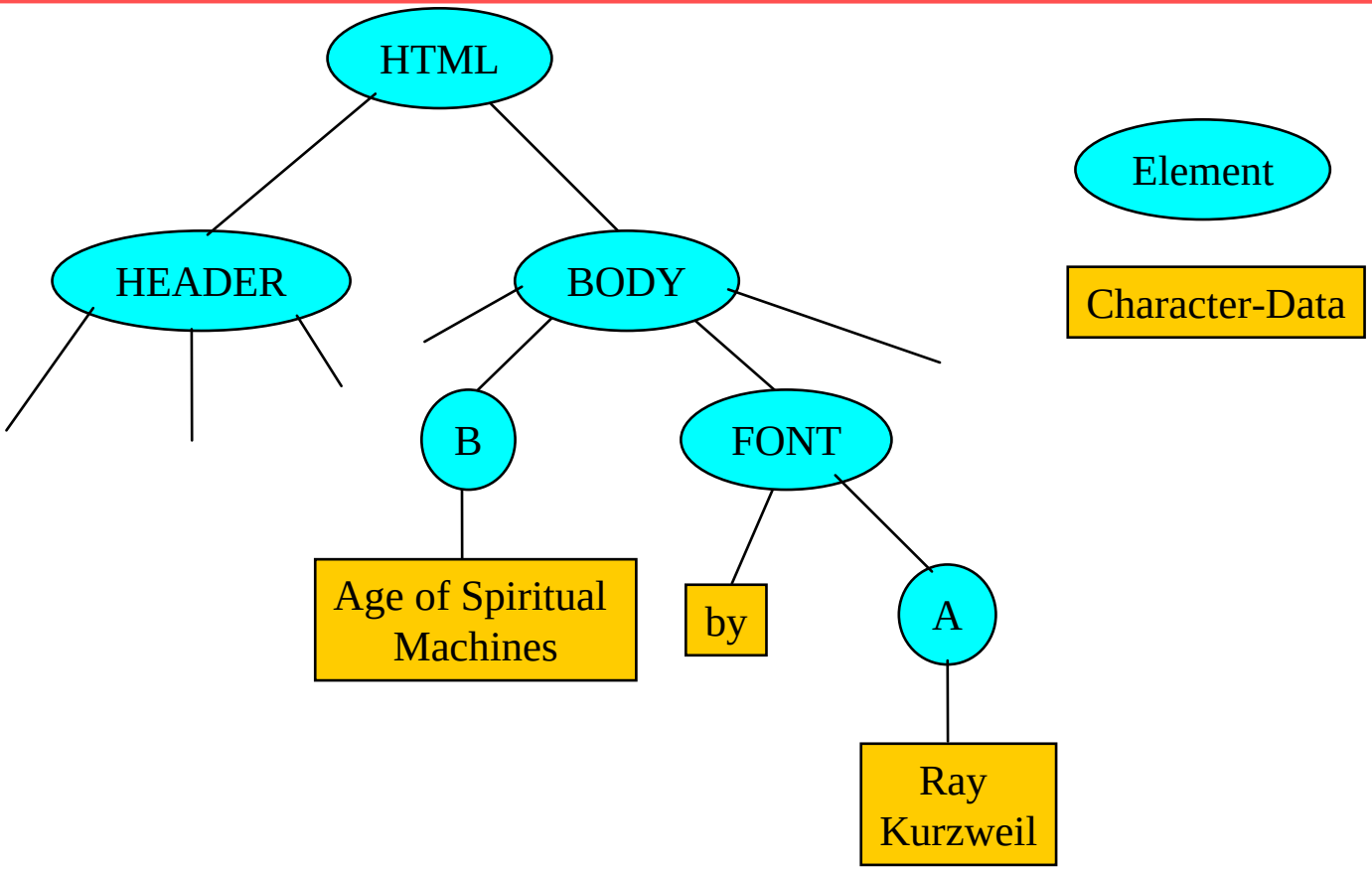
Web-Extraktion basierend auf DOM-Bäumen

- Web-Extraktion kann unterstützt werden, indem zunächst Webseiten geparst und als DOM-Bäume repräsentiert werden.

- Extraktionsmuster können dann als Wege von der Wurzel des DOM-Baums zum Knoten spezifiziert werden, der den zu extrahierenden Text enthält.

- Typischerweise greift man auf weitere Regex-Muster zurück, um den Teil im Zielknoten zu identifizieren, der extrahiert werden soll.

Beispiel DOM Baum Extraktion



Titel: HTML→BODY→B→CharacterData

Autor: HTML→ BODY→FONT→A→ CharacterData

Shop Bots

- Eine Anwendung der Web-Extraktion sind automatisierte Preisvergleichssysteme.
- Das System muss in der Lage sein, Informationen über Objekte (Produktspezifikation und Preise) aus einer Vielzahl von Webshops zu extrahieren.
- Der Anwender fragt eine einzige Seite an, die Informationen enthält, die von einer Vielzahl von Webshops extrahiert wurden und präsentiert dem Anwender umfassende Ergebnisse in einem einheitlichen Format, z.B. nach Preis gegliedert.
- Einige kommerzielle Systeme:
  - MySimon
  - Cnet
  - BookFinder

Informationsintegration

- Die Beantwortung mancher Fragen unter Benutzung des Webs erfordert integrierte Informationen von einer Vielzahl von Webseiten.
- Informationsintegration betrifft Methoden zur Automatisierung dieser Integrationsaufgabe.
- Erfordert Wrapper, um spezifische Informationen von Webseiten von spezifischen Sites exakt zu extrahieren.
- Behandle jede gewrappte Seite als Datenbank-Tabelle und beantworte komplexe Anfragen durch Verwendung einer Datenbank-Anfragesprache (z.B. SQL).

Informationsintegration: Beispiel

- Frage: Welches ist das nächste Kino von mir zuhause aus, das sowohl Monsters Inc. als auch Harry Potter spielt?
  - Extrahiere aus austin360.com Kinos und deren Adressen, wo Harry Potter und Monster's Inc. gespielt werden.
  - Überkreuze die beiden, um herauszufinden welche Kino beides spielen.
  - Frage mapquest.com nach Fahrtrouten von Deiner Privatadresse zur Adresse jedes dieser Kinos.
  - Extrahiere für jedes die Entfernung und die Fahrtanweisungen.
  - Sortiere die Ergebnisse nach der Fahrtentfernung.
  - Lege Fahrtanweisungen für das nächste Kino vor.