Informationsextraktion – Seminar

Vortrag über: Line Eikvil (1999) Information extraction from the World Wide Web

Referat Wintersemester 2002
Stephan Birkmann
November 2002

Übersicht

- Erläuterung der Problemstellung
- Abgrenzen des Themenbereichs
- Grundlagen
- Wrapper
- Kommerzielle Anwendungen
- Fazit

Definition

C. Cardie 1997:

An IE System takes as input a text and „summarizes“ it with respect to the user‘s domain of interest

Ziele der Informationsextraktion

- Wesentliche Informationen erkennen und in kompakter Form wiedergeben
- Anwendbarkeit auf
  - unbekannten Text
  - beliebig formatierte Textformen
- Also volles Textverständnis
- Endbenutzer Mensch

Probleme

- Wann wurde ein Text verstanden?
  - Volles Textverständnis (noch) nicht realisierbar,
  - deshalb Reduktion auf das Auffinden spezieller Informationen.
- Neuer Text bedeutet oft auch unbekannte Formatierung.
  - Benötigt also die Konstruktion eines neuen Patterns.
  - Es wird eine große Robustheit gegenüber strukturellen Veränderungen verlangt.

Themenabgrenzung

- Information Retrieval: Dokumentensuche aus Dokumentenmenge
- Information Filtering: Suche in nicht statischer Dokumentenmenge
- Textzusammenfassung: Schneller Überblick über den Inhalt
- Textkategorisierung: Selbständige Gruppierung von Texten
- Textklassifikation: Einordnen in vorgegebene Gruppen

Grundlagen

$$\text{Precision} = \frac{\text{korrekte Antworten}}{\text{produzierte Antworten}}$$

$$\text{Recall} = \frac{\text{korrekte Antworten}}{\text{mögliche korrekte Antworten}}$$

Grundlagen

Algorithmus / Mensch	Interessant	Nicht Interessant
Interessant	A	B
Nicht Interessant	C	D

$$\text{Precision} = \frac{A}{A + C} \qquad \text{Recall} = \frac{A}{A + B}$$

Textformen

- Frei: natürlichsprachlicher Text
  - syntaktische Beziehungen zwischen Wörtern
  - semantische Analyse
  - Geltungsraumerkennung von Namen
- Strukturiert: klar vordefinierte Formatierungsvorschriften
  - Extraktion durch Benutzen der Formatbeschreibung
- Semistrukturiert: ungrammatikalisch, telegrafisch

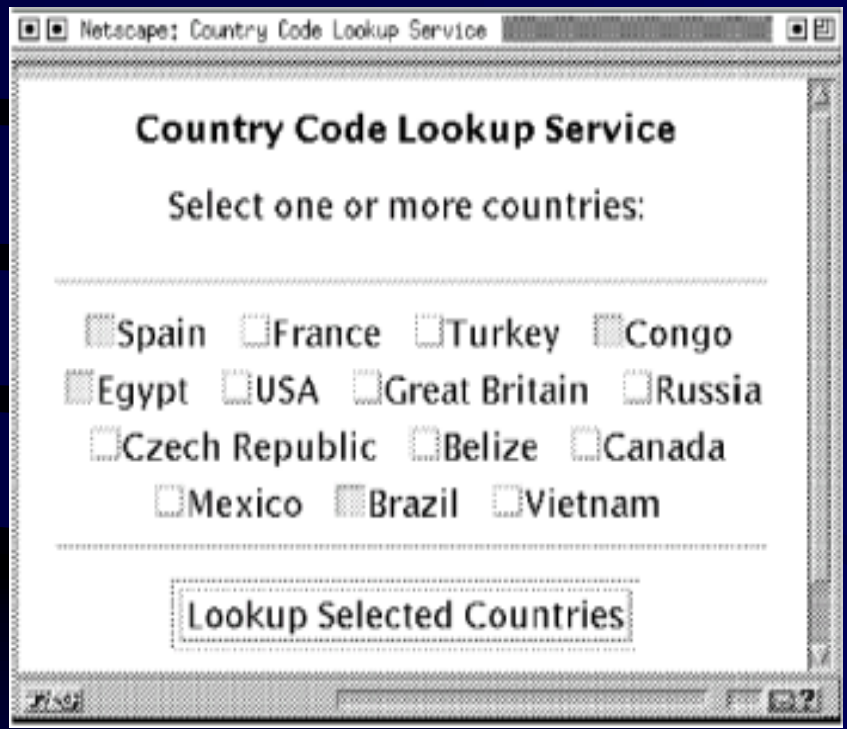
Internet

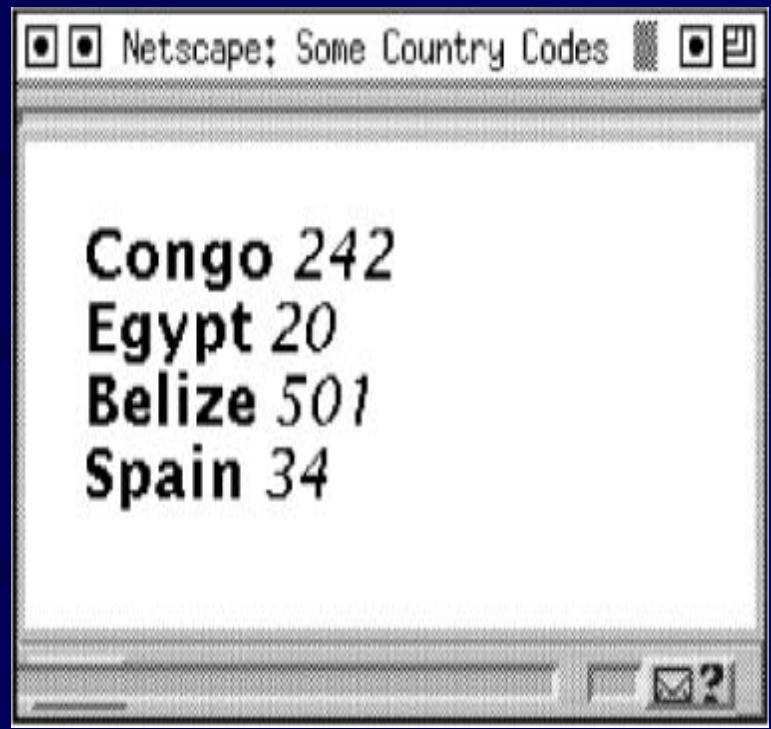
- Alle Textformen vorhanden
- Informationen werden oft erst auf Anfrage generiert (Hidden Web)
- Hyperlinks werden dynamisch von JavaScript erstellt
- Trotz HTML und XML keine Standards
- Global verteilte Informationen

Standardvorgehen

- Zerlegen und markieren
  - Zerlegen des Textes in einzelne Wörter (tokenising)
  - Bestimmen der Wortart (tagging)
  - Zuweisen der richtigen Wortart für jedes Wort
- Extraktion
  - Extraktionsregel wird mit vorliegendem Satz abgeglichen
  - Bei Erfolg wird die relevante Information identifiziert
- Ausgabe generieren
  - Die Information wird benutzt, um die vordefinierte Lücke im Fragebogen mit einer Antwort zu füllen

Trennsymbol-basiert





Trennsymbol-basiert

<HTML><TITLE>Some Country Codes</TITLE><BODY>
<B>Congo</B> <I>242</I>
<B>Egypt</B> <I>20</I>
<B>Belize</B> <I>501</I>
<B>Spain</B> <I>34</I>
</BODY></HTML>

$$\left\{\begin{array}{l}\langle\text{'Congo'},\text{'242'}\rangle,\\ \langle\text{'Egypt'},\text{'20'}\rangle,\\ \langle\text{'Belize'},\text{'501'}\rangle,\\ \langle\text{'Spain'},\text{'34'}\rangle\end{array}\right\}$$

Wrapper

- Tool zum gezielten Auffinden von Informationen
  - Muss an neue Formatierungen angepasst werden
  - Suche in unabhängigen Quellen benötigt also verschiedene Wrapper
  - Benutzt meistens nur Trennsymbolmuster
  - Entwickelt für die Abfrage von online generierten Seiten
- Bereitet gefundene Informationen zur Weiterverarbeitung auf.

Wrapper-Konstruktion

- Manuell: Programmierer analysiert die Grammatik
  - und programmiert den Wrapper
  - oder gibt die Grammatik in Wrappertool ein
- Halbautomatisch: Mensch zeigt dem Wrapper wo die Informationen zu finden sind
- Automatisch: ML-Techniken mit Hilfe von Beispieltexten

Inductive Learning

- Induktion: Aus gegebenen Grundbeispielen allgemeine Formel finden, aus der diese Beispiele (und andere) folgen.
- Zero-Order:
  - Attribute Value (Eigenschaft und ihr Wert)
  - Aussagenlogik
  - Kein Zusammenhang zwischen Objekten
- First-Order:
  - Prädikatenlogik erster Ordnung
  - Zusammenhang zwischen Objekten

Wrapper Induction

$$\left\{ \begin{array}{l} \langle\langle 50\ ,55\ \rangle, \langle 63\ ,66\ \rangle\rangle, \\ \langle\langle 78\ ,83\ \rangle, \langle 91\ ,93\ \rangle\rangle, \\ \langle\langle 105\ \ ,111\ \ \rangle, \langle 119\ \ ,122\ \ \rangle\rangle, \\ \langle\langle 134\ \ ,139\ \ \rangle, \langle 147\ \ ,149\ \ \rangle\rangle, \end{array} \right\}$$

The Wrapper Induction Problem

Benötigt: Ein Wrapper für die Quelle Q

Eingabe: Ein Satz von Beispielseiten S_1 aus Q
mit dem jeweiligen korrekten Label L

Ausgabe: Ein Wrapper W für die Quelle Q, so
daß W(S)=L für alle Beispielseiten

Test: Von Wrapper W auf Seitenmenge S_2 aus Q.
Ist $W(S_2) = L$?

LR Wrapper Class

procedure $\text{exec}_{\text{LR}}(\text{wrapper } \langle \ell_1, r_1, \ldots, \ell_K, r_K \rangle, \text{page } P)$

  $m \leftarrow 0$

  while there are more occurrences in P of ℓ_1 [i]

    $m \leftarrow m + 1$

    for each $\langle \ell_k, r_k \rangle \in \{\langle \ell_1, r_1 \rangle, \ldots, \langle \ell_K, r_K \rangle\}$

      scan in P to the next occurrence of ℓ_k; save position as $b_{m,k}$ [ii]

      scan in P to the next occurrence of r_k; save position as $e_{m,k}$ [iii]

  return label $\{\ldots, \langle \langle b_{m,1}, e_{m,1} \rangle, \ldots, \langle b_{m,K}, e_{m,K} \rangle \rangle, \ldots\}$

Eingabe und Ausgabe

<HTML><TITLE>Some Country Codes</TITLE><BODY>
<B>Congo</B> <I>242</I>
<B>Egypt</B> <I>20</I>
<B>Belize</B> <I>501</I>
<B>Spain</B> <I>34</I>
</BODY></HTML>

$$\left\{\begin{array}{l}\langle\langle 50\ ,55\ \rangle,\langle 63\ ,66\ \rangle\rangle,\\ \langle\langle 78\ ,83\ \rangle,\langle 91\ ,93\ \rangle\rangle,\\ \langle\langle 105\ ,111\ \rangle,\langle 119\ ,122\ \rangle\rangle,\\ \langle\langle 134\ ,139\ \rangle,\langle 147\ ,149\ \rangle\rangle,\end{array}\right\}$$

Wrapper-Klassen

- LR Left-Right
- HLRT Head-Left-Right-Tail
- OCLR Open-Close-Left-Right
- HOCLRT Head-Open-Close-Left-Right-Tail
- N-LR Nested-Left-Right
- N-HLRT Nested-Head-Left-Right-Tail

Kommerzielle Anwendungen

- Produktbeschreibung
- Restaurantführer
- Seminarankündigung
- Jobsuche
- Aktienmarktauswertung
- Krankenblattauswertung
- ...

Fazit

- Für die Praxis ist eine Zuverlässigkeit von 90% nötig
- Es werden weitaus portablere Systeme benötigt, um das gesamte Internet, auch Hidden Web, als eine Dokumentensammlung nutzen zu können
- Breite Anwendung wird die IE erst finden, wenn man sie mit anderen Techniken kombiniert
  - Information Retrieval (keine Angabe von Quellen)
  - Natural Language Generation (leicht lesbare Ausgaben)
  - Machine Translation (Sprachbarierre überwinden)
  - Data Mining (Analysen der zusammengestellten Information)