

OntoMiner

Intelligentes Datensammeln im Web

Stefan Kruppa

stefan.kruppa@informatik.uni-ulm.de

16. Juli 2004

A Einführung

Zielsetzung:

Ermöglichung des Semantic Web

Probleme:

- Daten im WWW größtenteils unstrukturiert
- Manuelle Strukturierung nicht möglich (Datenmenge)

→ Automatisierung nötig

→ OntoMiner

A Einführung

10-15 inhaltlich überlappende Websites

Strukturierte Daten



- 64News (B)
 - 88International
 - 70National
 - 72Washington
 - 74Business
 - 76Technology
 - 78Science
 - 80Health
 - 82Sports
 - 84New York Region
 - 86Education
 - 88Weather
 - 90Obituaries
 - 92NYT Front Page
 - 94Corrections
- 97Opinion (E)
 - 01Editorials/Op-Ed
 - 03Readers' Opinions
 - 09Advertisement
 - 11IMG
- 114Features (B)
 - 18Arts
 - 20Books
 - 22Movies
 - 24Travel
 - 26NYC Guide
 - 28Dining & Wine
 - 30Home & Garden
 - 32Fashion & Style
 - 34Crossword/Games
 - 36Cartoons
 - 38Magazine
 - 40Week in Review
 - 42Multimedia/Photos
 - 44Learning Network
- 1478services (B)
 - 51Archive
 - 53Classifieds
 - 55College
 - 57Book a Trip
 - 59Personals
 - 61Theater Tickets

Auslesen von Attribut-Wert-Paaren

z.B.: Gemüse = Kohlrabi
Gemüse = Kartoffel

Endergebnis:

- Baum, der Schlüsselbegriffe in Form einer Hierarchie ordnet
- Die Blattknoten entsprechen hierbei den Werten
- Die Nicht-Blattknoten entsprechen den Attributen

z.B.:

```
Wurzel
├── Gemüse
│   ├── Kohlrabi
│   └── Kartoffel
```

Ontologie:

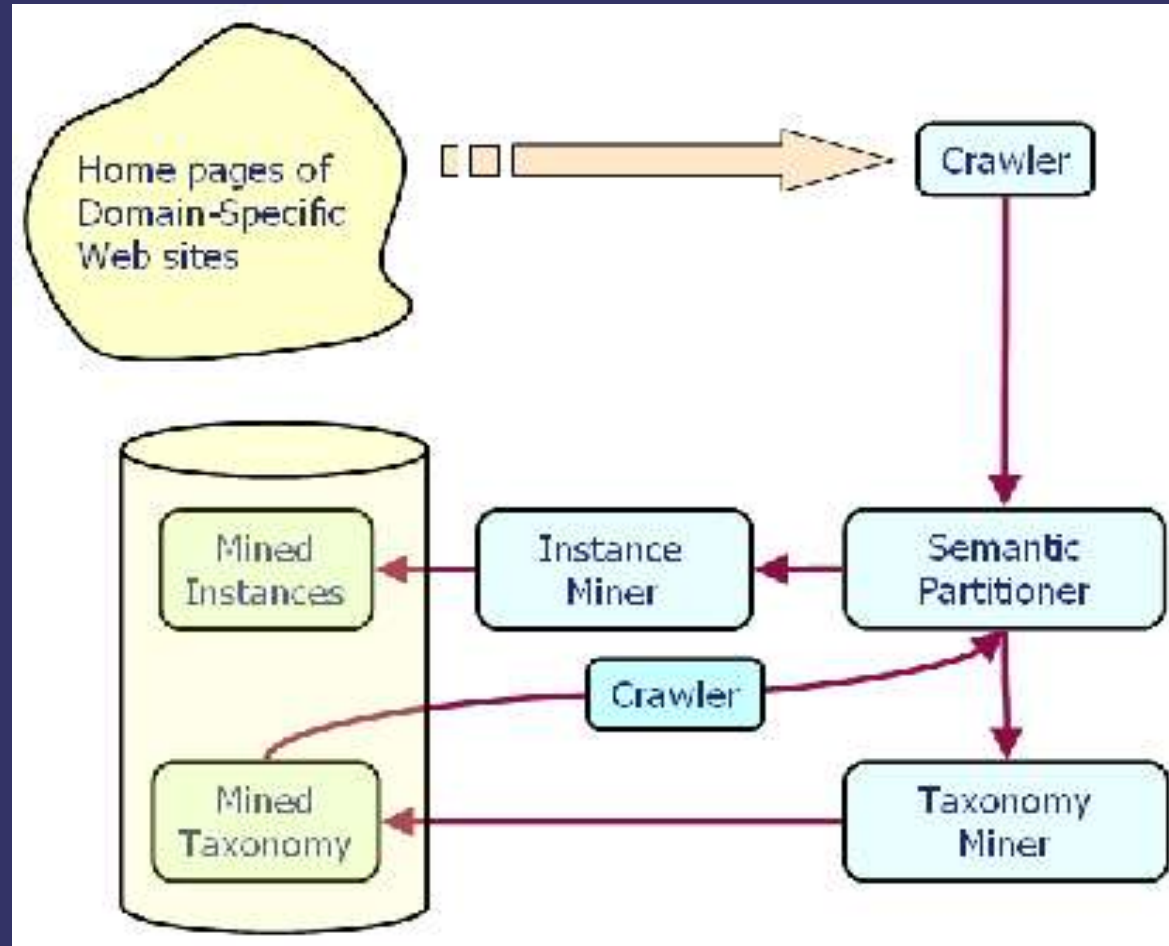
„Unter einer Ontologie versteht man in der Informatik im Bereich Künstliche Intelligenz ein formal definiertes System von Dingen und/oder Konzepten und Relationen zwischen diesen Dingen.“

„Eine Ontologie lässt sich vergleichen mit einer Datenbank - Struktur (Datenbankschema) und Inhalt (Daten) bilden ein Ganzes“

Taxonomie:

„Eine Systematik (auch Klassifikation, Taxonomie) ist eine planmäßige Darstellung von Klassen, Kategorien oder anderen Konzepten, welche nach bestimmten Ordnungsprinzipien gestaltet ist.“

Struktur von OntoMiner:



B Semantisches Partitionieren

ein Verfahren

XML-Baum → Semantikbaum

B Semantisches Partitionieren

Umwandlung einer einzelnen HTML-Seite in eine Sammlung von Begriffen, die den Inhalt der Seite charakterisieren.

- Einteilen einer Webseite in logische Segmente (Titelzeile, Navigationsleiste, Werbebanner usw.)
- Ermitteln von Datentabellen und Auslesen von Nutzinformatioren aus diesen
- Falls Datentabellen fehlen
 - ➔ komplizierter Weg (Hierarchisches Partitionieren):
 - HTML-Baum wird aufgelöst
 - Begriffe werden in neuen Baum geschrieben und dabei in Gruppen geordnet

C Taxonomy Mining

ein Verfahren

Semantikbäume → Taxonomie

Überblick Taxonomy Mining:

- Kernschritt des gesamten Prozesses
- Erstellt aus vielen Einzelbegriffen eine hierarchische Ordnung

Eingabe: Semantikbäume aus letztem Schritt (der für alle übergebenen Websites wiederholt wurde)

Ausgabe: fertige Strukturierung (Taxonomie) für das gewünschte Themengebiet

wie ? → dranbleiben :-)

Analyse der Semantikbäume

- Häufige Begriffe werden als wichtig erachtet und extrahiert
- Begriffe, die den häufigen strukturell ähnlich sind, werden extrahiert
- Überflüssige Begriffe werden wieder gelöscht
- Gruppieren der Begriffe in **Konzepte**
- Beseitigung von Redundanz
- Herausarbeiten von Eltern-Kind-Beziehungen aus dem semantisch partitionierten Baum → Taxonomie
- Wiederholen des Verfahrens auf alle Unter-Seiten
- Erweitern der Taxonomie in die Tiefe

D Instanzenextraktion

ein Verfahren

Füllen der Taxonomie

Finden von Instanzen zu den Konzepten

- Vergleich zweier repräsentativer Unter-Seiten einer Website
- Diejenigen Segmente, die Informationen über Instanzen enthalten, werden anhand von syntaktischen Regelmäßigkeiten gefunden
- Diese Regelmäßigkeiten werden ausgenutzt, um die Instanzinformationen aus all den Unter-Seiten auszulesen, die diese Strukturen zeigen

E Zusammenfassung

E Zusammenfassung

Eingeben von 10-15 inhaltlich überlappenden Websites



Umwandeln in Semantikbäume (Sortierung des Inhalts)



Erstellen der Taxonomie (Konzepte)



Erzeugen von Untertaxonomien (für Unterbegriffe wiederholen)



Füllen der Taxonomie

F Bewertung

~~OntoMiner = KI~~

aber: gute Vorarbeiten für das Semantic Web

G Literatur

- H. Davulcu, S. Vadrevu, S. Nagarajan.
OntoMiner: Bootstrapping Ontologies From
Overlapping Domain Specific Web sites.
IEEE Intelligent Systems no.5, Sept./Okt. 2003, S.24-33
- Wikipedia Online-Enzyklopädie, <http://de.wikipedia.org>

Tschööö :-)