

# Invisible Web

## Scatter/Gather-Clustering für semistrukturierte Daten

Praxisprojekt des Studienganges Angewandte Informatik im Wintersemester 2003/04

Norbert Fuhr, Gudrun Fischer

Einführung:

# Hintergrund und Aufgabe

Gudrun Fischer

Blockseminar in Hagen, 17. Oktober 2003

# Invisible Web

- Web-Seiten, die für „normale“ Suchmaschinen „unsichtbar“ sind
  - Hinter Portalen
  - In Datenbanken
  - Erst auf Anfrage generiert
  - Nicht-indexierte Formate
- Synonyme:
  - Deep Web, Hidden Web

# Probleme

## ■ Finden

- Web-Directories für das Invisible Web
- Portale

## ■ Zugriff

- Anfragen formulieren (aber wie?)

## ■ Übersicht

- Inhalt einer Invisible-Web-Quelle?
- Browsing?

# Yahoo für das Invisible Web

## ■ Yahoo

- Hierarchische Kategorien
- Statisch
- Manuelle Pflege

## ■ Invisible Web

- Hierarchische Inhaltsübersicht
- Dynamisch
- Automatisch

⇒ **Scatter/Gather**

# Scatter/Gather-Clustering: Ablauf

- Aufteilung der Datenmenge  $M$  in Cluster
- Anwender wählt  $x$  Cluster aus
- Daten aus diesen Clustern werden neue Datenmenge  $M$
- Da capo

# Scatter/Gather-Clustering: Eigenschaften

- Hierarchisch
- Dynamisch, auf Anfrage
- Individuell pro Anwender

- Details:

André Nurzenski und Tang Zhihong

Sonntag, 19.10.2003

# Semistrukturierte Daten

- Mehr Struktur als Text
- Weniger Struktur als DB-Datensätze
- Paradebeispiel: XML
- Herausforderungen:
  - Verschiedene Schemata
  - Extraktion
  - Ähnlichkeits-/Abstandsmaße
  - Darstellung

# Aufgabe

- Scatter/Gather-Werkzeug
- Semi-strukturierte Daten
- Austauschbare Komponenten:
  - Ähnlichkeits-/Abstandsmaß
  - Clustering-Algorithmus
  - Inhaltliche Zusammenfassung für Cluster
  - GUI
  - möglicherweise weitere...

# Nächste Schritte

- Vorbereitung
  - Name
  - Praktische Fertigkeiten
  - Blockseminar: Hintergrundwissen
- Vision
  - Wie soll es aussehen?
- Modellierung
  - Komponenten
  - Interaktion
- Unit-Tests