Der MLHTBrowser:

Interaktive Exploration von Kollektionen auf verschiedenen Abstraktionsebenen

Gudrun Fischer, Michael Chojnacki
Universität Duisburg-Essen

6. Workshop des GI-Arbeitskreises Knowledge Discovery
Karlsruhe, 01.03.2005

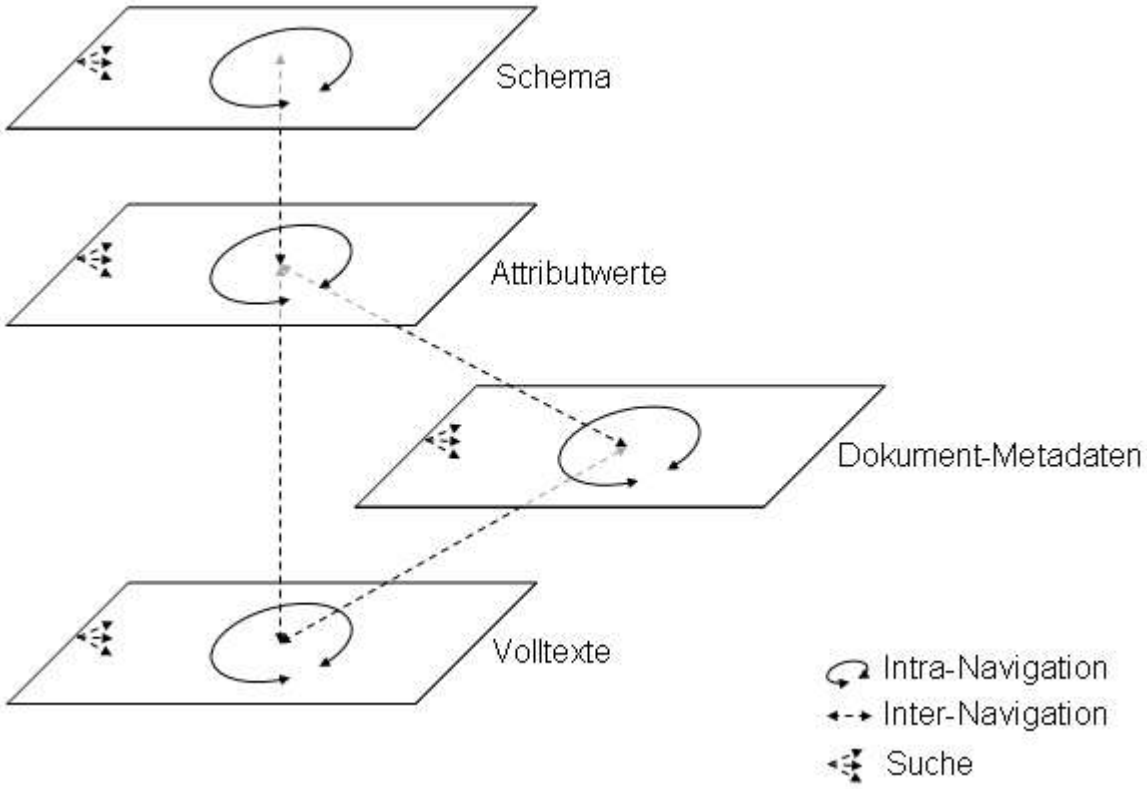
Motivation

- Unbekannte, neue Kollektion
  z.B. aus dem Deep Web
- Semistrukturierte Daten
  XML, HTML, Metadaten – mehr als nur Text
- Sofortiges Erforschen
  interaktiv, flexibel

Cluster-basiertes Browsing

- Struktur durch Clustering
  - statisch, offline erzeugt
    z.B. in Portalen
  - dynamisch, online berechnet
    z.B. für Suchergebnisse
- Scatter/Gather
  - Auswahl einer Teilmenge
  - Neugruppieren
  - flexiblere Navigation

Abstraktionsebenen: Multi-Level-Hypertext



Browsing in MLHT: Beispiele

- Browsing innerhalb einer Ebene
  - von Metadaten zu Metadaten (z.B. CiteSeer)
  - in einer Klassifikationsstruktur (z.B. ACM)
- Browsing zwischen Ebenen
  - von Metadaten zu Dokumenten (z.B. CiteSeer)
  - von Klassifikation zu Dokumentmetadaten und zurück (z.B. ACM)
  - von Autoren zu Dokumentmetadaten und zurück (z.B. ACM)

Semistrukturierte Daten

- Strukturell homogene Kollektion: Feld = Pfad
- Beispiele:
  - in Volltexten: Kapitel, Abschnitt, Titel ...
  - in Metadaten: Titel, Erscheinungsjahr, Verlag ...
    (Feld = Attribut)
- Aspekte einer Ebene:
  - Betrachtung eines Feldes der Ebene oder
  - der Inhalte aller Felder einer Ebene
- Datentypen

Beispiel: Aspekte in bibliographischen Metadaten

- Auf der Metadatenebene:
  - Titel
  - Abstract
  - alle Felder als Text
- Auf der Attributwertebene:
  - Autoren
  - Schlüsselwörter
  - Verlage
  - Erscheinungsjahre
  - Werte aller Attribute als Text (Terme)

MLHT: Verallgemeinertes Konzept

- Level
  - Abstraktionsebenen
- Aspekte/Teilansichten
  - Projektion auf Feld(er)
- Datentypen
  - Text, Zahlen, Namen, Terme ...
- Anordnungsmöglichkeiten je nach Datentyp
  - Liste, Intervalle, Cluster

Berrypicking: mehr Interaktion - mehr Information

- Aufsammeln interessanter Objekte
  - Implizites Relevanzurteil
- Dynamische Relevanzschätzung
  - Ähnlichkeit angezeigter Objekte zu Korbinhalt
- Visualisierung
  - Markierung (Farbe, Balken, Schattierung...)
  - Anordnung nach Ranking

MLHTBrowser

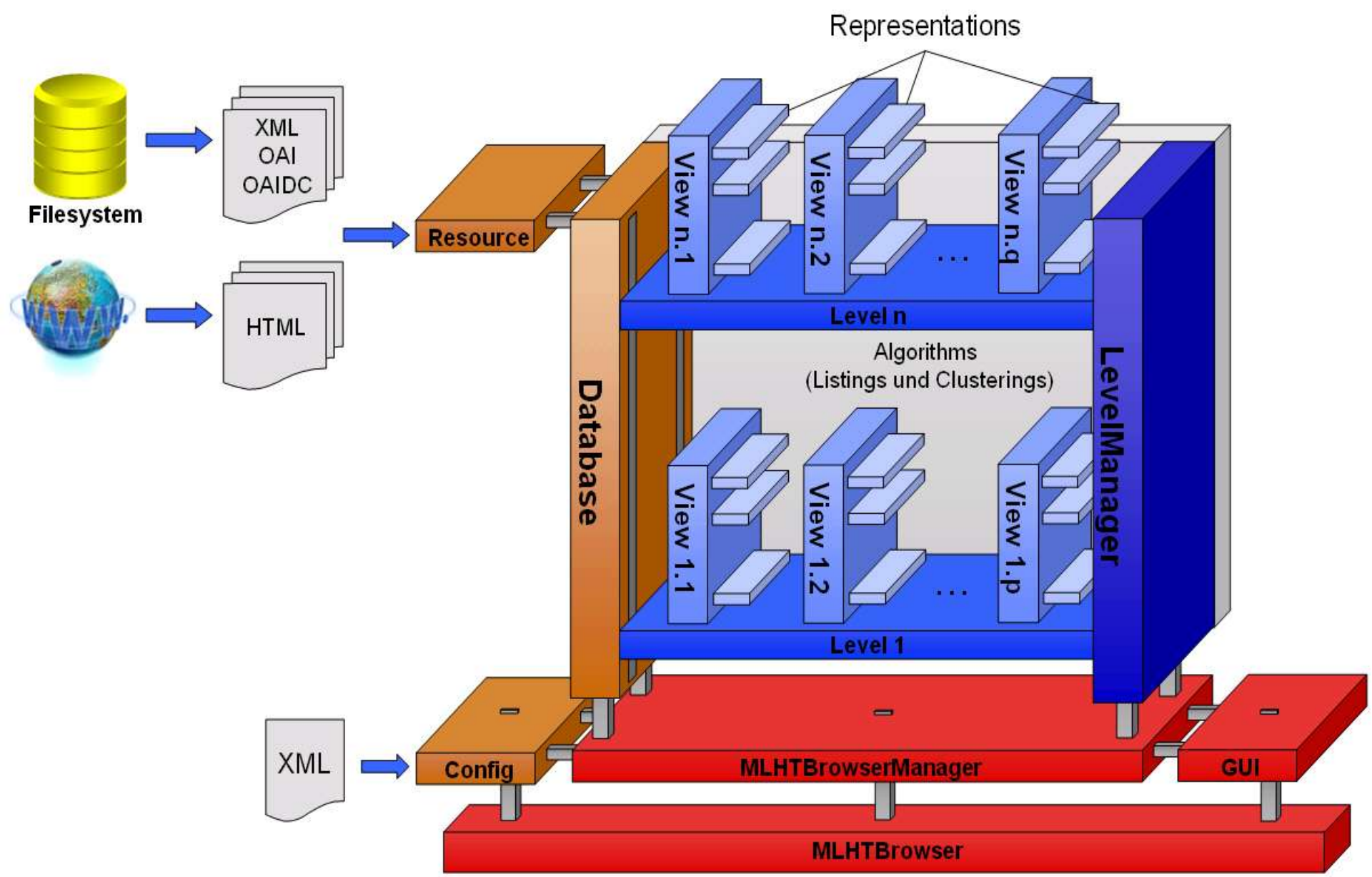
- Scatter/Gather-Clustering auf Metadaten- und Attributwertebene
- Erweiterter Scatter-Schritt
  - Wechsel der Anordnung
  - Wechsel des Aspekts
  - Wechsel der Ebene
- Berücksichtigung von Datentypen
- Berrypicking und dynamische Relevanzschätzung
  - Visualisierung durch Ampelfarben

Demo...

- 520 Einträge (Metadaten) aus der CompuScience-Kollektion (FIZ Karlsruhe)
- Aspekte der Attributwertebene:
  - Autoren
  - Erscheinungsjahre
  - alle Terme
- Aspekte der Metadatenebene:
  - ganze Datensätze

Implementierung: Framework

- Konfigurierbar
  - Datenextraktion und -modellierung
  - Dokument- und Termgewichte
  - Clusteralgorithmen
  - verfügbare Darstellungen
  - verfügbare Abstraktionsebenen
- Erweiterbar
  - andere Clusteralgorithmen
  - andere Datentypen (Ähnlichkeitsmaße, Anordnung)


Representations
Filesystem
XML
OAI
OAIDC
HTML
Resource
View n.1
View n.2
View n.q
Level n
Algorithms
(Listings und Clusterings)
Database
LevelManager
View 1.1
View 1.2
View 1.p
Level 1
XML
Config
MLHTBrowserManager
GUI
MLHTBrowser

Implementierung: Clustering

- Ähnlichkeitsmaße
  - Dokumente: Kosinusmaß über Termvektoren (verschiedene Termgewichte möglich)
  - Terme: gemeinsames Auftreten in Dokumenten (verschiedene Gewichtungen möglich)
  - Attributwerte: Ähnlichkeit der Dokumente, in denen die Attributwerte vorkommen
- Dimensionalität: Feature-Begrenzung
- Algorithmen: Varianten von Buckshot

Anwendungen

- Standalone
  - für schnelle Exploration, ohne Vorbereitung
- GoogleBrowser
  - als Frontend einer komplexeren Google-Suche
- In Daffodil
  - zum Clustern und Browsen von Suchergebnissen in einer digitalen Bibliothek
- In Pepper
  - als GUI für Browsing in P2P-Netzen

Zusammenfassung

- Modell für Browsing in Multi-Level-Hypertext
- Hochinteraktives Werkzeug
- Konzeptionelle und funktionale Integration von:
  - Browsing in Multi-Level-Hypertext
  - Scatter/Gather-Prinzip
  - Felder und Datentypen
  - Berrypicking und dynamische Relevanzschätzung
- Vielseitig konfigurierbar

Ausblick

- Verbesserte Darstellung der Cluster
- Probabilistisches Clustering
  - für die Metadaten- und Dokumentenebene
  - Kombination von Feldähnlichkeiten
- Verbesserte Relevanzabschätzungen
  - Fusion der Körbe
  - Ähnlichkeit verschiedenartiger Objekte
- Möglichkeiten von Vorprozessierung
- Clustering mit externem Wissen
- Weitere Datentypen

Danke

für Ihre Aufmerksamkeit!