Terminologieorientiertes Website-Matching mit COMA++

Vortrag zur Diplomarbeit

Diplomand: Sebastian Stoll Betreuer: Andreas Thor

Gliederung

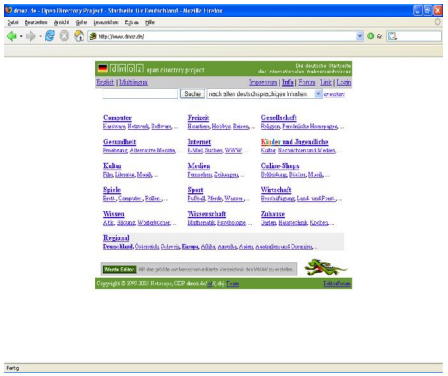
- Einleitung
- Ansatz der Diplomarbeit
- Realisierung des Ansatzes
- Evaluierung
- Zusammenfassung
- anschließend Diskussion

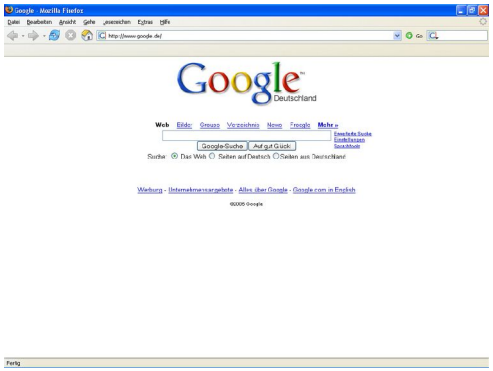
1.Einleitung

Problematik

Gängige Methoden zum Finden von Inhalten

1. 2. 3.





Browsing

- zeitaufwändig
- Verlinkung der Seiten?

Web-Directories

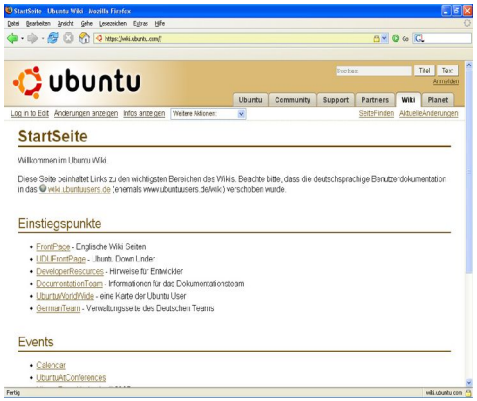
- handverlesen
- nicht immer aktuell
- zeitaufwändig

Suchmaschinen

- Formulierung der Suchanfragen bereitet Probleme
- Ranking

Recommendation Services

- Empfehlungen für den Nutzer
- Verteilte Ansätze (z.B. ALEXA Web Search)
- Auswertung von Informationen über den Nutzer (z.B. URL)



Knopfdruck

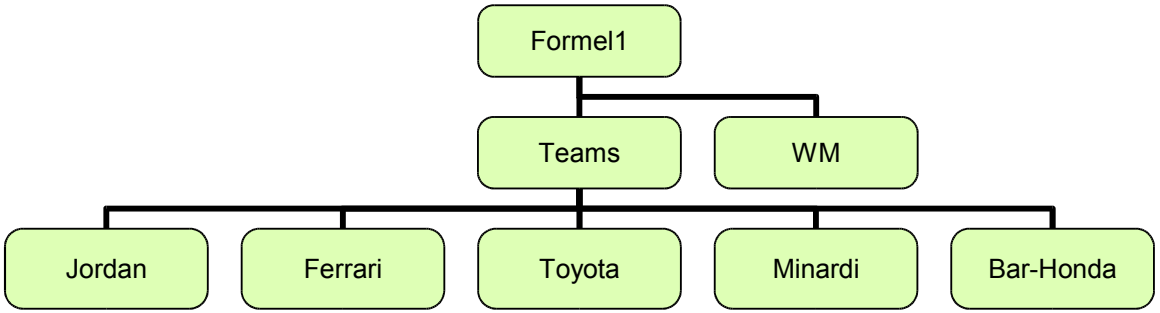
Ausgangsseite

Empfehlungen

Frage: Ermittlung von semantisch ähnlichen Seiten?

2.Ansatz der Diplomarbeit

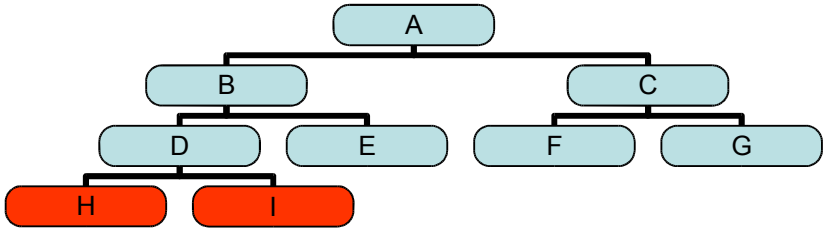
Strukturierter Index für semantisch ähnliche Webseiten



- strukturierter Index als Datenbasis für Recom - mendation Services
- Index dargestellt über Ontologie
- Ontologie charakterisiert eine Domäne

Motivation für Verwendung eines strukturierten Index

- Steuerung der Granularität bzw. der Sicht auf die Domäne
- Erweiterung der Suchergebnisse
  - semantischer Abstand zweier Klassen
  - Schwellenwert
- Einbeziehen von in der Ontologie enthaltenen Informationen
  - z.B. Restriktionen



$$sim(k1,k2) = \left(\frac{2 \cdot Tiefe(k_{spez})}{Tiefe_{k_{spez}}(k1) + Tiefe_{k_{spez}}(k1)} \right)$$

$$sim(H,I) = \left(\frac{2 \cdot 2}{3+3} \right) = 2/3$$

Terminologieorientierte Ontologien

- Terminologie (Fachwortschatz): „Gesamtbestand der Begriffe und ihrer Benennungen in einem Fachgebiet“
- Repräsentation der Semantik einer Klasse durch weitere Terme als den Klassennamen notwendig
- Auszeichnung der Klassen einer Ontologie mit Begriffen aus domänspezifischer Terminologie
- Ausschluss von Mehrdeutigkeit oder Synonymie

Beispiel: Auszeichnung der Klasse Jordan durch die Terme Jordan, Trevor Carlin, Tiago Monteiro, Narain Kathikeyan

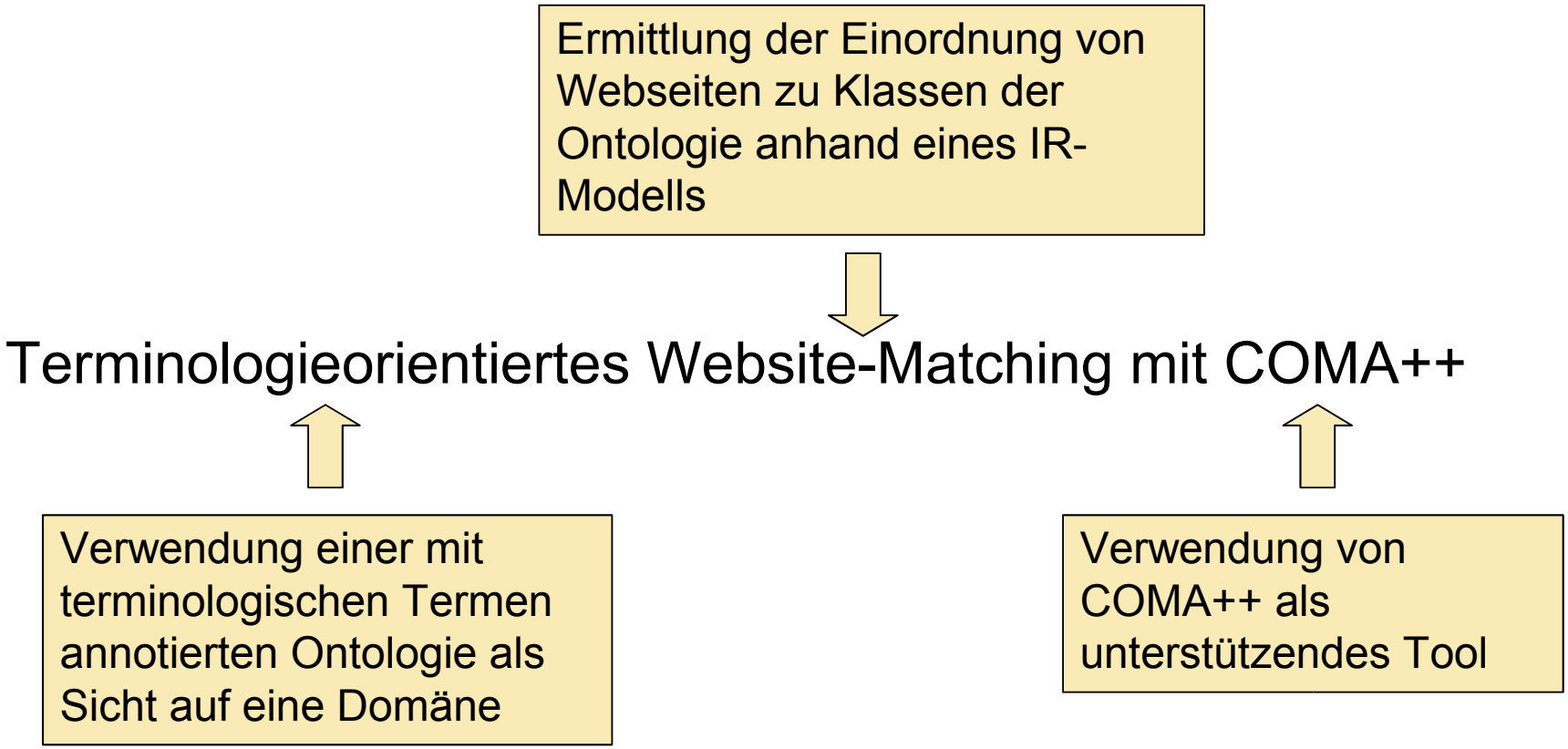
Website-Matching

- Ermittlung der Ähnlichkeit zweier Webseiten bzw. von Ontologie und Webseite
- Ähnlichkeitsberechnung auf Basis eines IR-Modells
- Berechnungsgrundlage für IR-Modell ist interne Darstellung
- Output ist eine Zuordnung der Webseiten zu den Klassen der Ontologie → strukturierter Index

Verwendung von COMA++

- Unterstützung der Ontologieerstellung
- Tool zur Evaluierung
  - Verwendung von intendierten Mappings
  - Kombination von Matchergebnissen
  - Hilfe bei Auswahl eines Ähnlichkeitsmaßes
- Visualisierung der Webseitenstruktur
- Fragmentbasiertes Matching

Zusammenfassung des Ansatzes

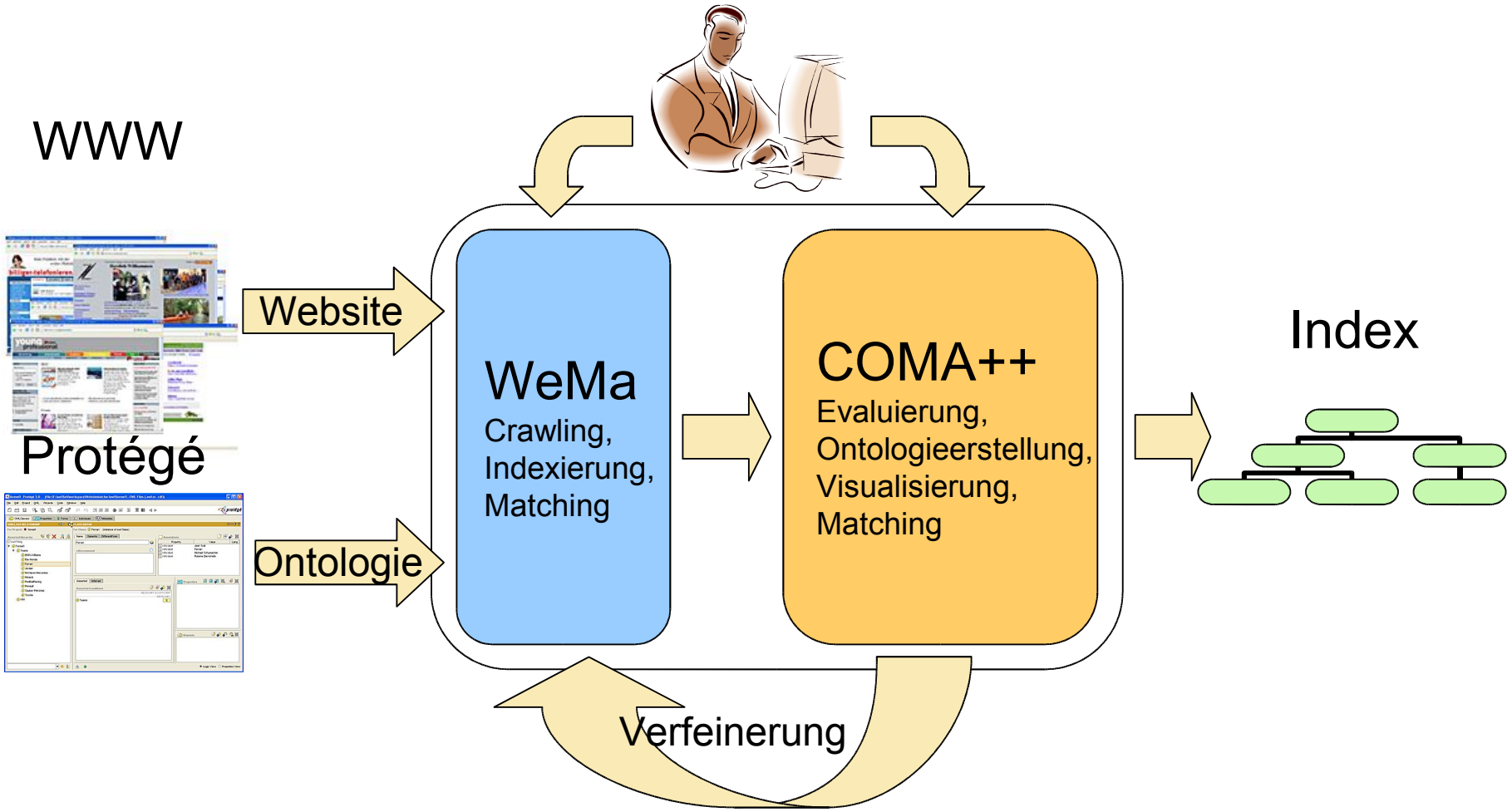


3.Realisierung des Ansatzes

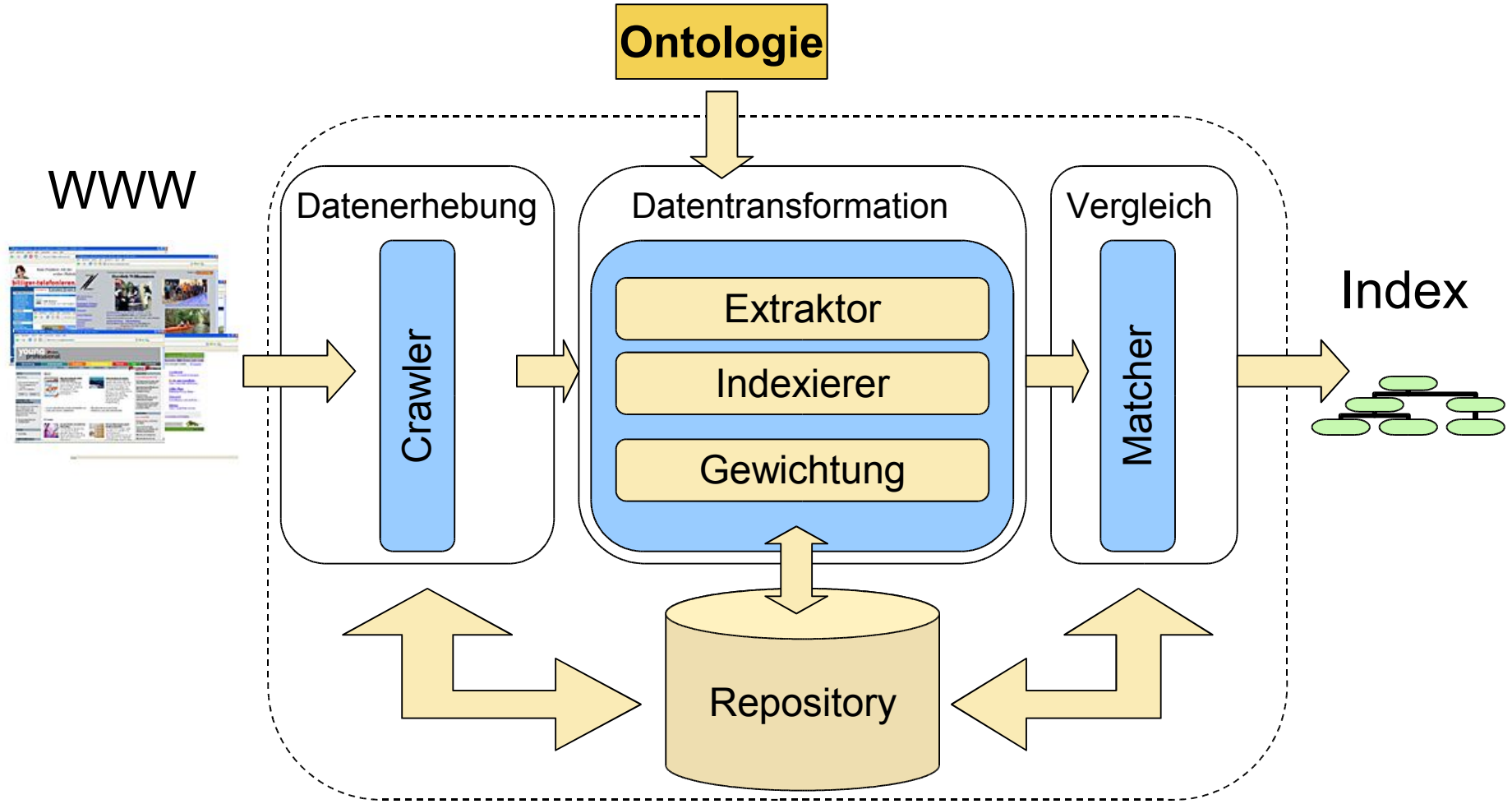
Umfang der Implementierung

- WebsiteMatcher – WeMa
  - Prototyp eines WIRS
  - Crawling
  - Indexierung
    - Datenbereinigung
    - Datentransformation
  - Matching
  - Ontologieimport
- Integration in COMA++
  - Datenstrukturen
  - Matching

Anwendersicht



Entwicklersicht WeMa

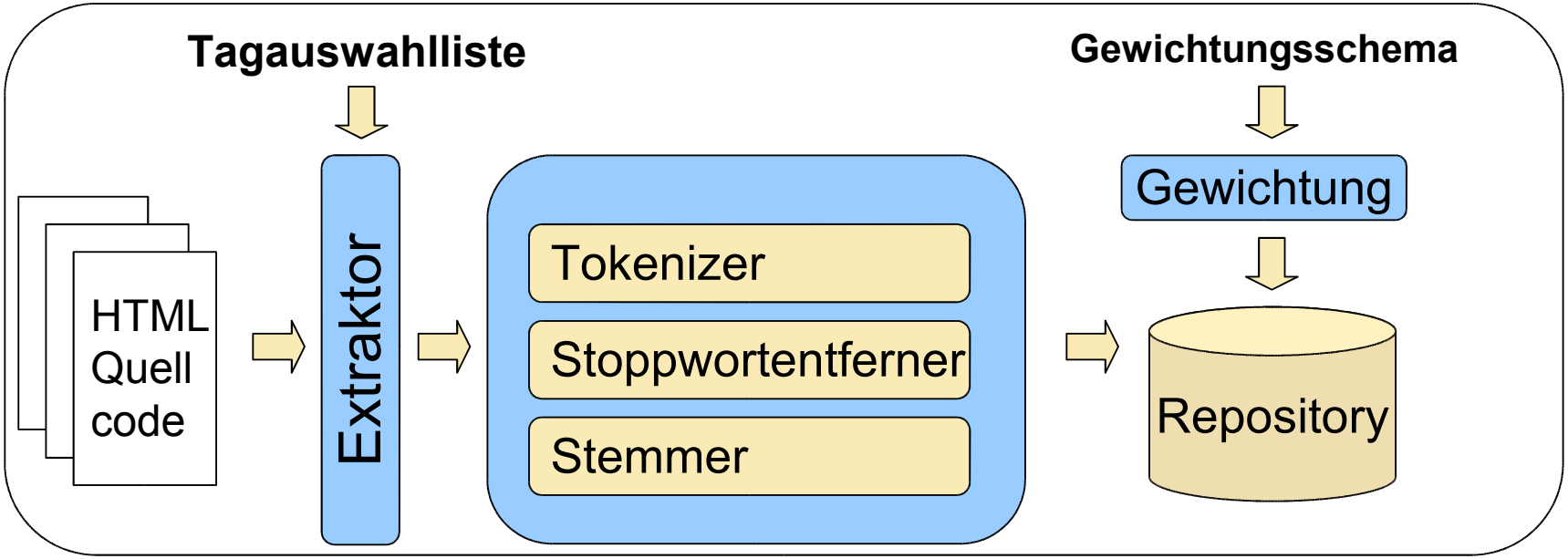


Datenerhebung+Datenhaltung

- Crawler basierend auf Hyperspider von David Aumüller
  - Erstellung eines Strukturbaums mit „minimal click path “ Strategie
  - Möglichkeit der grafischen Darstellung und Umwandlung in andere Formate
- Speicherung aller WeMa relevanten Daten in einem Repository
- Generisches Datenmodell
  - gleiche Repräsentation von Webseiten und Ontologien
  - basiert auf der Idee der invertierten Liste

Indexierung

•Ziel: Transformation in interne Darstellung sowie Erhöhung der Datenqualität



•Extraktoren für die Selektion von Inhalten

•Stoppwortentfernung und Stemming für Verbesserung der Datenqualität

•Vektorraummodell (VRM) mit TF·IDF Maß für Gewichtung

•zusätzlich Gewichtung von Strukturmerkmalen

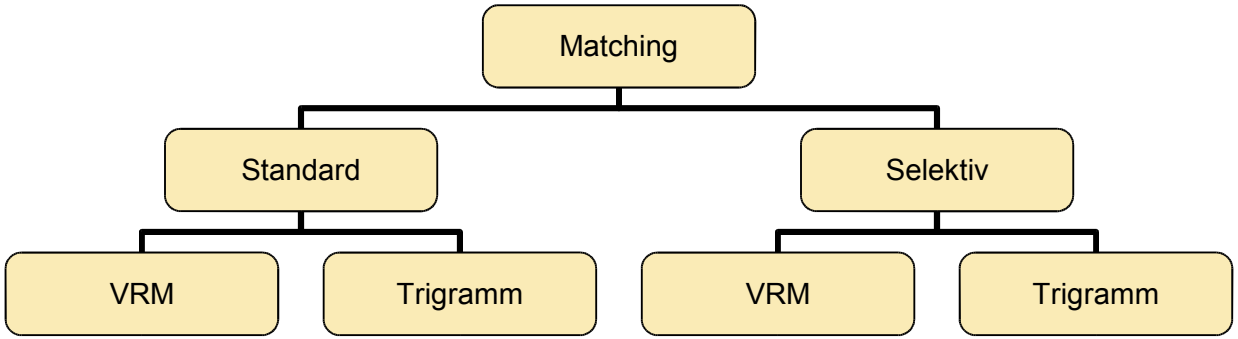
OWL zur Darstellung der Ontologien

- Standard des W3C
- Annotation der Ontologien über „label“- und „comment“-Tags
- ebenfalls Darstellung im VRM
- gleiche Abfolge der Indexierung wie bei Webseiten
- Erstellung mit Hilfe von Protégé

```
<owl:Class rdf:ID="Jordan">
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Teams"/>
  </rdfs:subClassOf>
  <rdfs:label>Jordan</rdfs:label>
  <rdfs:label>Trevor Carlin</rdfs:label>
  <rdfs:label>Tiago Monteiro
  </rdfs:label>
  <rdfs:label>Narain Karthikeyan
  </rdfs:label>
 </owl:Class>
```

Matching

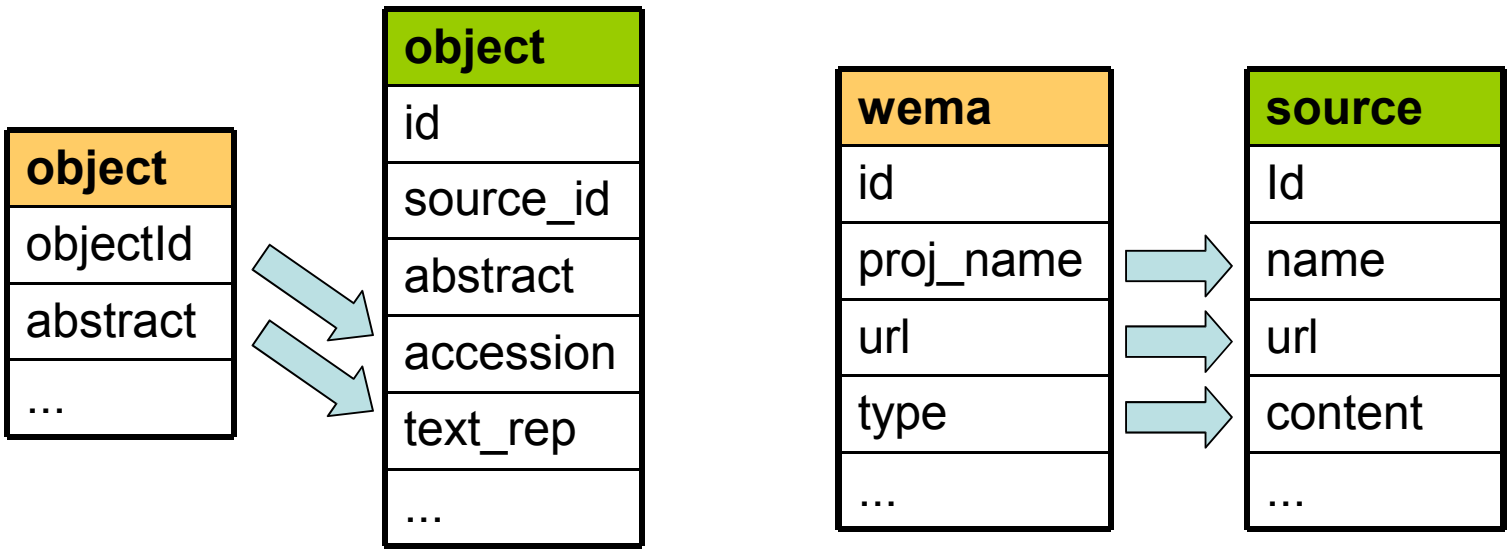
- Instanzbasiertes und Ontologiebasiertes Matching
- Komplexität der Ähnlichkeitsberechnung O(n·m)
- Selektion der zu vergleichenden Komponenten über eine Tagauswahlliste
- Optimierung der Berechnung über Vorauswahl der Vergleichskandidaten



Maße sowohl für instanzbasiertes als auch ontologiebasiertes Matching

Integration in COMA++(1)

- Ziel war die Erweiterung von COMA++ um die Funktionalität des Website – Matchings
- Datenimport erfolgte über ID – Mapping von WeMa nach COMA++



Integration in COMA++(2)

- Erstellen einfacher Ähnlichkeitsmaße: SIM_DOC_VECTOR und SIM_DOC_TRIGRAMM
- Integration der optimierten Ähnlichkeitsberechnung
- Auswahl über das GUI

Name	Tagauswahl	Maß
Complete	Alle Tags	VRM, Trigramm
Title	TITLE	VRM, Trigramm
Metatag	DESCRIPTION,KEYWORDS	VRM, Trigramm
Meta+Title	TITLE, DESCRIPTION, KEYWORDS	VRM, Trigramm
Body	BODY	VRM, Trigramm

4.Evaluierung

Ziel der Evaluierung

- Messung des Retrievalverhaltens
- Untersuchung des Einflusses der einbezogenen Informationen
- Ermittlung des besten Matchers
- Identifizierung von Stärken und Schwächen des Ansatzes
- Aufzeigen von Verbesserungsmöglichkeiten

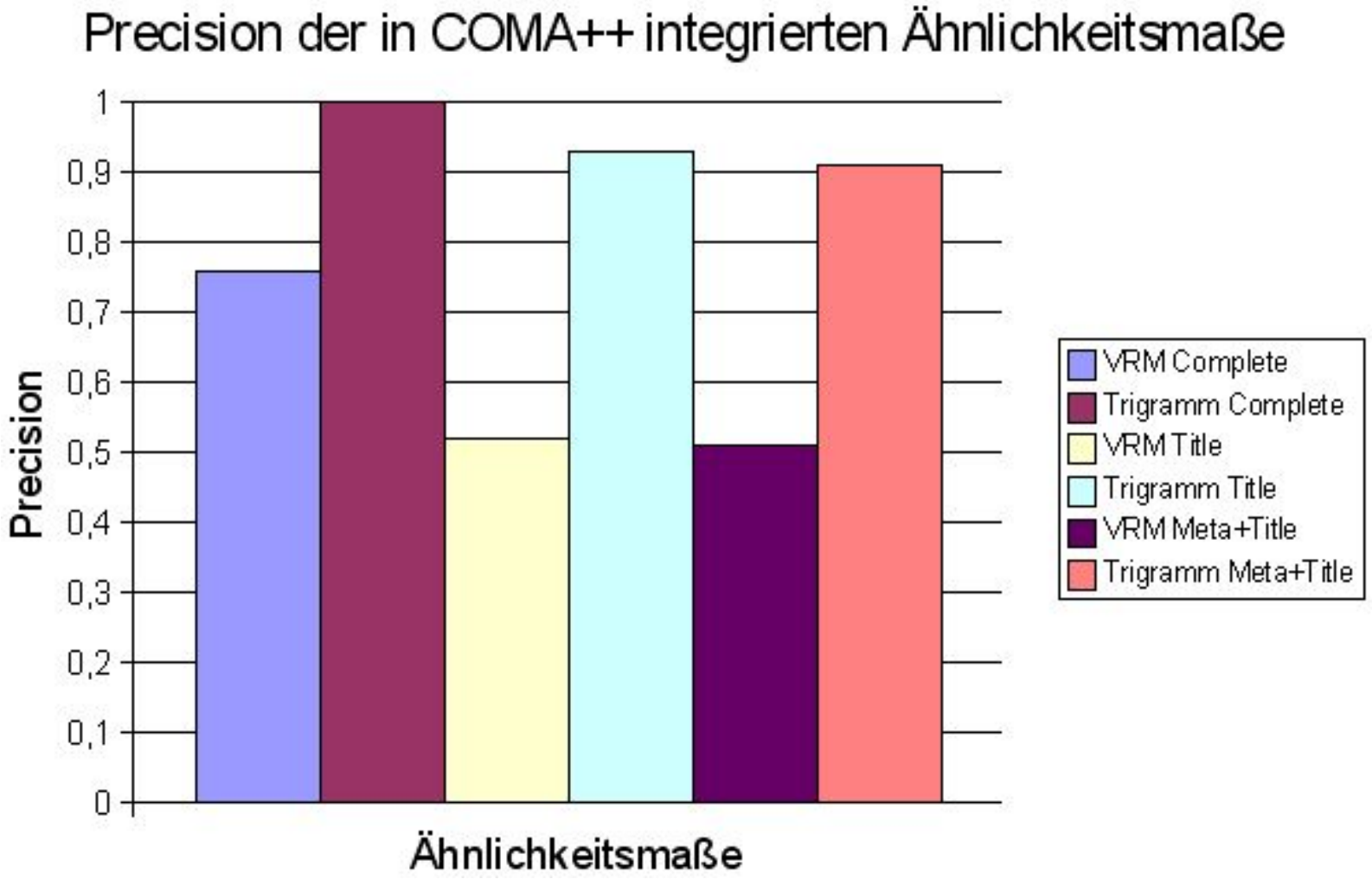
Anwendungsfall Personal News

- Präsentation von aktuellen Nachrichten in Abhängigkeit der Interessen des Nutzers
- Darstellung der Interessen über Ontologie
- Grundlage der Datenerhebung waren 13 Internetauftritte deutscher Nachrichtenmagazine mit Beschränkung auf Formel1-Inhalte
- Ermittlung der durchschnittlichen Precision
- Evaluierung mit COMA++

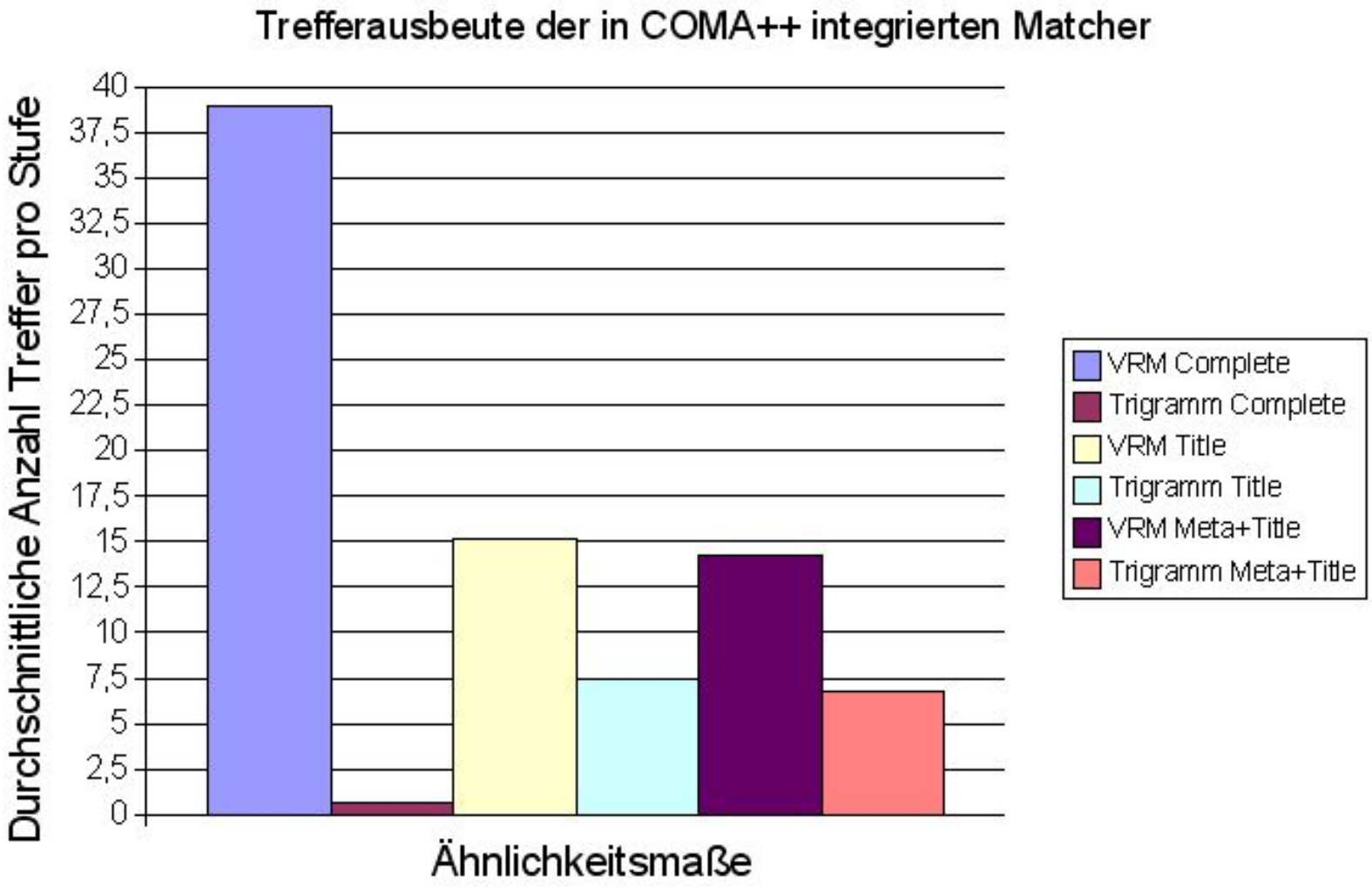
Parameter:

- Wema
  - Crawling bis Stufe 4
  - Extraktion von HTML und Metatags
  - Stoppwortentfernung und Stammformreduktion
- COMA++
  - Matcher „Complete“, „Title“, „Meta+Title“
  - Schwellenwerte {0.2,0.3,...,0.7}
  - VRM und Trigrammmaß

Auswertung Personal News (1)



Auswertung Personal News (2)



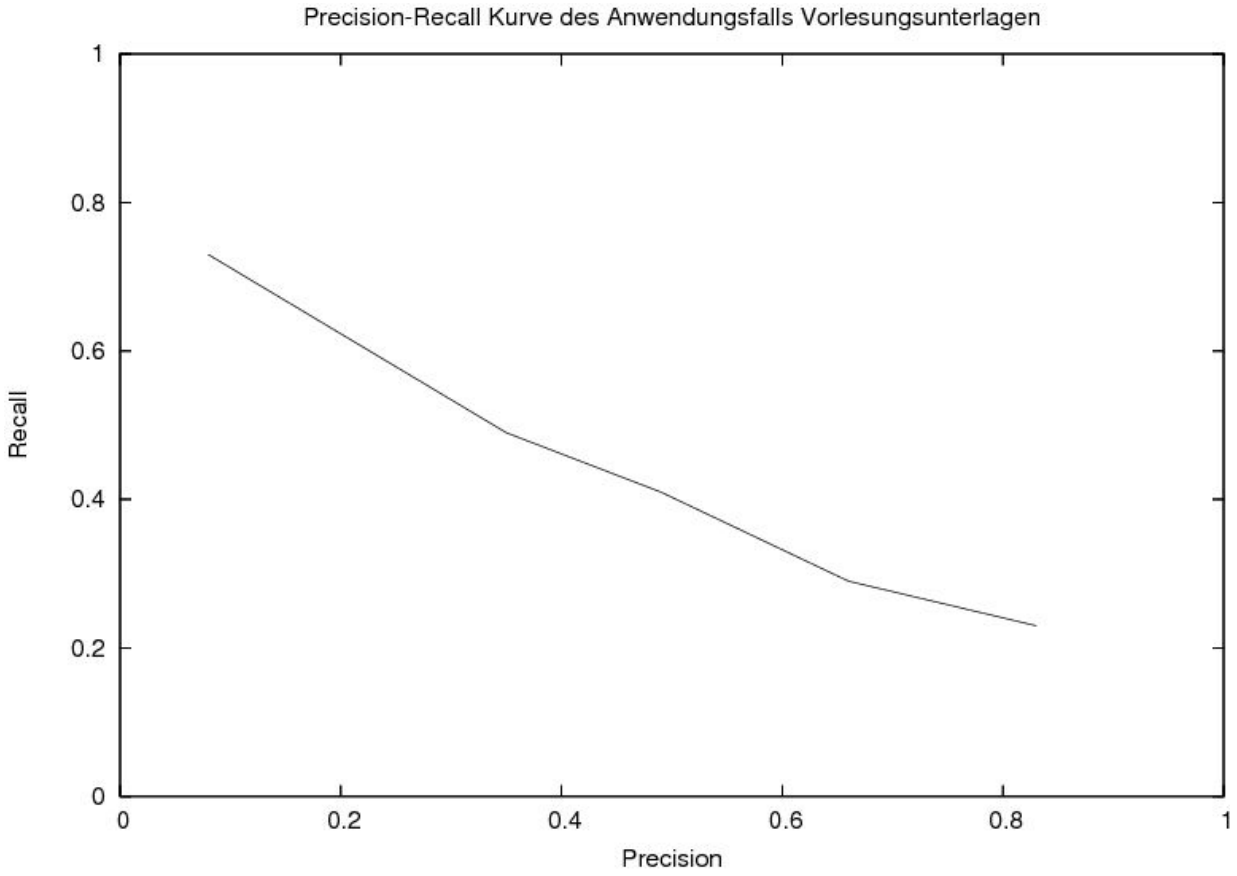
Anwendungsfall Vorlesungsunterlagen

- Recommendation Service für das Auffinden verwandter Inhalte
- Darstellung der Themengebiete Datenbanken und Informationssysteme über eine Ontologie
- Grundlage der Datenerhebung waren 10 Webseiten von Professuren dt. Universitäten
- Ermittlung der durchschnittlichen Precision+Recall
- Evaluierung in COMA++ mit Hilfe eines intendierten Mappings

Parameter:

- Wema
  - Crawling bis Stufe 4
  - Extraktion von HTML Tags
  - Stoppwortentfernung und Stammformreduktion
- COMA++
  - Matcher „Complete“
  - Schwellenwerte {0.5,0.7, 0.9}
  - VRM und Trigrammmaß

Anwendungsfall Vorlesungsunterlagen Auswertung



Durchschnittliche Precision: **0.56**

Durchschnittlicher Recall: **0.26**

Typische Fehlerkonstellation

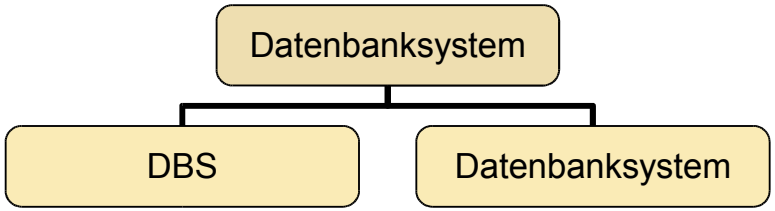
- Eigennamen:
  - „Lauryn Williams“ vs. „Frank Williams“
- Homonyme in Titel oder Überschriften:
  - „Der Jordan als Kloake“ vs. „Team Jordan“
- fehlende Diskriminanz:
  - „Datenbanksysteme“ oder „Transaktionssysteme“ in Kombination mit Unabhängigkeit der Terme
- weitere: Metaphern, unterschiedliche Schreibweisen und Abkürzungen

- Datenerhebung
  - Hyperspider im Entwicklungsstadium
  - Erhöhung der Performanz durch Multithreading und Datenbankunterstützung
  - Automatisierung der Webseitenauswahl über focused Crawling oder Verwendung von Web Directories

- Erweiterung Datenvorverarbeitung
  - Behebung des Eigennamenproblems durch Part-of-Speech Tagging
  - Verbesserung der Stoppworteliminierung durch Tagging bzw. Unabhängigkeit

Der	Sieg	von	Fernando	Alonso	war	großartig.
ART	**N**	**PREP**	**NE**	**NE**	**VAFIN**	**ADJA**

- Verwendung eines Thesaurus und damit Reduzierung von Synonymen, Abkürzungen und Schreibweisen auf einen Term bzw. ein Konzept



Betrachtungen zum Gesamtsystem und Erweiterungsmöglichkeiten(3)

- Ontologieerstellung
  - sehr zeitaufwändig, setzt Domänenkenntnisse voraus
  - Problematik spezifischer Domänenontologien,
- Erweiterungsmöglichkeiten der Ontologieerstellung
  - Automatisierung → „Onto Learn“ – System
  - Annotation über „Topic Distillation“, ähnlich Google News

- Matching
  - hängt von zwei Faktoren ab: interner Darstellung der Webseiten und der Ontologie
  - bestes Ähnlichkeitsmaß ist der Complete-Matcher
  - Um Problematik der Unabhängigkeit von Termen zu begegnen, können Term – Term Korrelationen einbezogen werden
  - Einbeziehung von Abstandsinformationen

Zusammenfassung

- Konzeption und Realisierung eines Systems zur Erstellung einer Datenbasis für RS
- gute Ergebnisse wenn Terme diskriminierend sind (Eigennamen bei Formel1)
- mittelmäßige Ergebnisse bei stark spezialisierten Domänen (Vorlesungsunterlagen)
- sinnvoller Einsatz von COMA++ zur Evaluierung, Ontologieerstellung, Vorarbeiten und Matching
- Vielfalt des WWW bedingt Erweiterungen zur Verbesserung der Retrieval-Qualität
- Automatisierung des Ontologieentwurfes und der Annotation von Webseiten wäre hilfreich für den Ansatz