Prof. Dr. Matthias Hemmje

Kurs 01874

Informations- und Wissensmanagement im Internet

LESEPROBE

Fakultät für
Mathematik und Informatik

FernUniversität in Hagen

Inhalt

Kurseinheit 1 – Einführung, semantische Integration und Informationsmanagement

Kurseinheit 2 – Ontologiesprachen für das Internet, Erzeugung von Ontologien und Metadaten

Kurseinheit 3 – Retrieval und Integration

Kurseinheit 4 – Statistische und räumliche Informationen

Einleitung

Dieser Text beinhaltet die erste Kurseinheit des Kurses 1874 „Informations- und Wissensmanagement im Internet“. Ziel dieses Kurses ist es, aufbauend auf dem Kurs 1873 „Daten- und Dokumentenmanagement im Internet“, weiterführende Themen zu den Grundlagen der Wissensrepräsentation zu behandeln. Die adäquate Repräsentation von Wissen und die hierfür verwendeten Formate hängen dabei in starkem Maße von der Art des Wissens und von der Anwendung des repräsentierten Wissens ab. Wissen kann sich auf Personen, Fakten, Vorgehensweisen, Regeln, Gesetzmäßigkeiten etc. beziehen und umfasst sowohl statische als auch dynamische Aspekte.

Ontologien als wesentliche Mittel zur Konzeptionalisierung des Wissens über eine Domäne gewinnen aktuell im Rahmen der Entwicklung des Semantic Web als Grundlage der interoperablen Repräsentation an Bedeutung. Der Kurs beschäftigt sich dementsprechend sowohl mit der Grundidee der Ontologien als auch mit Formaten zu deren Repräsentation. Dabei wird insbesondere auf die aktuell entstehenden Standards zur adäquaten Repräsentation von Ontologien im Web (DAML+OIL, OWL) eingegangen und diese werden zu RDF in Beziehung gesetzt.

Um das Wissen von Informationen für den Benutzer anschließend auch wieder auffindbar zu machen, wird in Kurseinheit 3 eine Einführung in das Information Retrieval gegeben. Es werden hierzu unterschiedliche Modelle vorgestellt, die auch aktuell im WWW zum Einsatz kommen.

Die vorliegende erste Kurseinheit stellt nach einer kurzen Motivation und Einführung zunächst anhand einer relevanten Auswahl der aktuell vorliegenden Ergebnisse der wissenschaftlichen Forschung und Literatur das Themenfeld „Semantische Integration“ vor. Dabei baut der Kurstext auf Inhalten einer Seminarveranstaltung auf, die an der FernUniversität in Hagen im WS 2006/7 durchgeführt worden ist.

1 Motivation und Einführung

Das Internet ist eine gigantische Sammlung von Informationen aller Art. Von Kochrezepten bis zu Bauplänen kann man hier fast jede Information finden - wenn man genug Geduld und Zeit aufbringt. Die pure Anzahl von Web-Sites wird für den Nutzer zu einem Problem. Es kommt zu einem Informationsüberfluss. So meldete z. B. Yahoo im August 2005 eine Indexgröße von 19,2 Milliarden Seiten. [1] Dabei ist ein erheblicher und meist der qualitativ

hochwertigere Teil der Informationen noch nicht einmal zugänglich, da er auf Anfrage dynamisch aus Datenbeständen zusammengesucht und auf dynamisch generierten Seiten (vgl. Deep Web) angezeigt wird. Darüber hinaus sind vielfältige Informationen derzeit in privaten Datenbanken, etwa von Unternehmen oder Behörden, in digitalen Bibliotheken oder Intranets der breiten Nutzung entzogen. Diese Datenmengen in einem anderen als dem ursprünglich vorgesehenen Kontext oder in Kombination anwenden zu können, böte verschiedenste Möglichkeiten, davon zu profitieren.

Deep Web

Die adäquate Repräsentation von Information und Wissen bildet daher eine wichtige Grundlage für deren effektive Erfassung, Verwaltung, Vermittlung und Nutzung. Die Verwendung geeigneter Standards erleichtert dabei den Austausch von Information und Wissen zwischen Menschen, zwischen Menschen und Maschinen sowie zwischen Maschinen. Solche Standards stellen ein gemeinsames Grundverständnis unabhängig von einer individuellen Aushandlung von Formaten sicher und ermöglichen damit auch dynamische und wechselnde Kooperationsbeziehungen, die in unserer durch das Internet global vernetzten und sich kontinuierlich veränderten Welt zunehmend an Bedeutung gewinnen.

1.1 Das semantische Web als Fortentwicklung des bestehenden Netzes

Die Verfügbarkeit einer kritischen Masse an Information, deren Integration zum einen einen echten Mehrwert darstellt und zum anderen die Unübersehbarkeit und Unübersichtlichkeit der vorhandenen Informationsquellen, begründet eine zunehmende Bedeutung von Diensten, die eine Art „integrierten Zugriff“ anbieten. Die Bereitstellung eines solchen Zugriffs auf die beschriebene Menge von heterogenen, autonomen und verteilten Informationsquellen und die automatische Verarbeitbarkeit der vorgefundenen Informationen ist somit Voraussetzung für Web-Services aller Art sowie für fortgeschrittene Anwendungen in e-commerce-, e-banking-, e-learning- und anderen Bereichen. Diese Bereitstellung wird allgemein als gemeinsame Informationsnutzung (*information sharing*) bezeichnet.

Information sharing

Derzeit wird der größte Teil des Web-Inhalts im Hinblick auf menschliche Nutzung bereit- und dargestellt. Der Nutzer muss die vorgefundene Information sichten, bewerten und ggf. selbständig mit Informationen aus anderen Quellen verknüpfen. Die Aktivitäten der Benutzer werden, abgesehen von verschiedenen Suchmaschinen, kaum von Programmen unterstützt. Die Benutzer stehen daher vor folgenden Problemen [2]:

- Es stellt sich das Problem einer großen Anzahl von Suchergebnissen (Treffer) bei geringer Präzision/Relevanz dar [3].

- Im Gegensatz dazu ergeben sich keine oder nur wenige Treffer.
- Die Suchergebnisse sind stark vom Vokabular abhängig. Wenn die Suchbegriffe nicht mit den Schlüsselwörtern der relevanten Dokumente übereinstimmen, werden diese nicht gefunden, obwohl sie in ihrer Bedeutung übereinstimmen.
- Es werden nur einzelne Web-Seiten als Suchergebnis geliefert. Dadurch kann es notwendig werden, weitere Anfragen zu stellen. Danach müssen die Ergebnisse quasi manuell und damit mit erheblichem kognitiven Aufwand verknüpft werden.
- Auch bei erfolgreicher Suche müssen die Benutzer in den gelieferten Dokumenten die interessierende Information manuell und damit mit erheblichem kognitiven Aufwand herausfiltern.
- Suchergebnisse sind für andere Programme kaum nutzbar und müssen manuell integriert werden.

Maschinelle Verarbeitung

Um diese Probleme bearbeiten zu können, ist die grundlegende Möglichkeit der maschinellen Verarbeitung aller im Web verfügbaren Inhalte notwendig. Dabei muss die dargestellte Information nicht unbedingt immer im Wortsinne vollständig oder auch nur teilweise vom Rechner interpretierbar werden, sondern es reicht in vielen Fällen aus, sie derart maschinell verarbeitbar zu machen, dass mithilfe verschiedener Software-Tools das gewünschte Ergebnis geliefert wird.

Semantisches Netz

Dieses Konzept findet als so genanntes Semantisches Netz (*semantic web*) aktuell Verwendung. Computerbasierte Dienste bieten dabei dem Nutzer bei Suche, Sichtung und Bewertung von Informationen Unterstützung an, indem sie den semantischen Hintergrund der dargestellten Informationen interpretieren und automatisch maschinell weiterverarbeiten. Darüber hinaus sollen Beziehungen zwischen den interpretierten Informationen vom System erkannt werden. Methoden der Wissensrepräsentation und -verarbeitung sollen sowohl eine gemeinsame Nutzung als auch Zugriff auf verteilte Informationsquellen ermöglichen.

1.2 Probleme bei der Etablierung eines Semantischen Netzes

Semantische und syntaktische Heterogenität

Als zentrales Problem erweist sich im Semantischen Netz die semantische und syntaktische Heterogenität der Informationsquellen und die mangelnde Struktur der Information auf Web-Seiten. Um eine sinnvolle Nutzung der

heterogen Informationsquellen gewährleisten zu können, ist die automatische Zusammenführung ihrer Inhalte unumgänglich. Dies wird unter dem Begriff der Semantischen Integration (*semantic integration*) behandelt. Die technische Umsetzung wird in der Regel durch das Mediatorenmodell (vgl. 2.5.1) realisiert.

Semantische Integration

Die eindeutige Darstellung der Semantik der Information in schwach strukturierter Umgebung stellt sich dabei als eine wesentliche Voraussetzung dar, um die mit der Heterogenität der Daten verbundenen Probleme zu lösen [4]. Probleme, die sich aus der Heterogenität der Daten ergeben, sind aus dem Bereich der verteilten Datenbanksysteme bereits gut bekannt.

Es lassen sich dabei drei wesentliche Problemfelder benennen:

- die Syntax – z. B. heterogene Datenformate
- die Struktur – z. B. Homonyme, Synonyme oder unterschiedliche Attributnamen in Datenbanktabellen
- die Semantik – z. B. die beabsichtigte Bedeutung von Ausdrücken in speziellem Zusammenhang oder Anwendung

1.3 Lösungsansätze

Während zu den ersten beiden Punkten bereits ausdifferenzierte Lösungen aus der Datenbank-Forschung existieren, befinden sich die maschinelle Interpretation und die Verarbeitung der Semantik noch in einem frühen Forschungsstadium. Deshalb soll im Folgenden vor allem dieser Aspekt besondere Darstellung erfahren.

Dabei stellt das semantische Web kein neues System parallel zum bestehenden World Wide Web dar, sondern es soll sich sukzessiv aus diesem herausbilden. Die Entwicklung verläuft dabei auf vier Ebenen:

1. die Ebene der technischen Aspekte eines Computernetzwerks
2. die Ebene des Webs als Benutzungsschnittstelle für die Interaktion zwischen Mensch und Internet
3. Wissensebene
4. Anwenderebene eines Wissensnetzwerks [30] im Sinne einer Basis für soziale Netzwerke

Auf der dritten Ebene kann das Semantische Web als verteilte Datenbank bzw. Wissensbasis betrachtet werden, deren Inhalt in maschinenlesbarer und

-verarbeitbarer Weise vorliegt und von Softwarewerkzeugen zur Informations-Verarbeitung und zum -Management automatisch verarbeitet wird. Zudem werden den Benutzern bestimmte Aufgaben abgenommen:

- Informationssuche
- Informationssammlung
- Informationsklassifikation
- Informationsfilterung
- Informationsmanagement
- Informationsmining
- Informationsentdeckung
- Informationsbewertung

Hierzu ist die Ergänzung des aktuellen Inhalts der Web-Dokumente um Daten über deren Semantik, so genannte Metadaten, notwendig. Dabei ist die fortgeschrittene Standardisierung der Darstellung verschiedener Inhalte auf syntaktischer Ebene von Vorteil. Als Beispiele für solche Standards können Datenbank-Schnittstellen wie Open Database Connectivity (ODBC) oder W3C Web-Technologien wie HTML, XML oder RDF (vgl. Kurs 1873) genannt werden, die die Integration unterschiedlicher Informationsquellen erlauben. Deren Potential soll für die Integration auf semantischer Ebene genutzt werden.

2 Semantische Integration: Konflikte und Lösungsansätze

2.1 Heterogenität auf struktureller und semantischer Ebene

Die Integration heterogener Datenquellen erfordert mehr als nur das Darstellen der Daten in einer gemeinsamen Syntax. Die oben beschriebenen syntaktischen Standards ermöglichen zwar die einheitliche Abbildung und Strukturierung von Informationen im Web, wodurch die automatische Verarbeitung sowohl von lokal vorliegender Information als auch aus entfernten Quellen stammender Information erheblich erleichtert wird. Die unterschiedliche Struktur und Semantik von Informationen aus unterschiedlichen Quellen, wie sie im Web vorzufinden sind, führen jedoch zu differierenden Problemen, die das Eingreifen auf mehreren Ebenen erforderlich

machen [5]. Syntaktische Homogenität ist also notwendige, aber nicht hinreichende Bedingung für die gemeinsame Informationsnutzung. Konflikte treten dabei auf folgenden Ebenen auf:

- Datenmodellebene
- Datenschemaebene
- Dateninstanzebene

2.2 Heterogenitätskonflikte zwischen Datenmodellen

Datenquellen können sich deutlich in der Darstellung von Daten unterscheiden (z. B. als Tabellen, Objekte, Dateien usw.). Hiermit wird die syntaktische Ebene angesprochen. Das Abgleichen heterogener Datenquellen macht ein gemeinsames Daten-Modell erforderlich, worauf die Informationen aus den unterschiedlichen Quellen abgebildet werden können. Es müssen geeignete Transformationen gefunden werden, um die Daten in das entsprechende Modell zu überführen. (XSLT, XQuery [siehe Kurs 01873]). Abbildung 2.1 soll dies veranschaulichen:

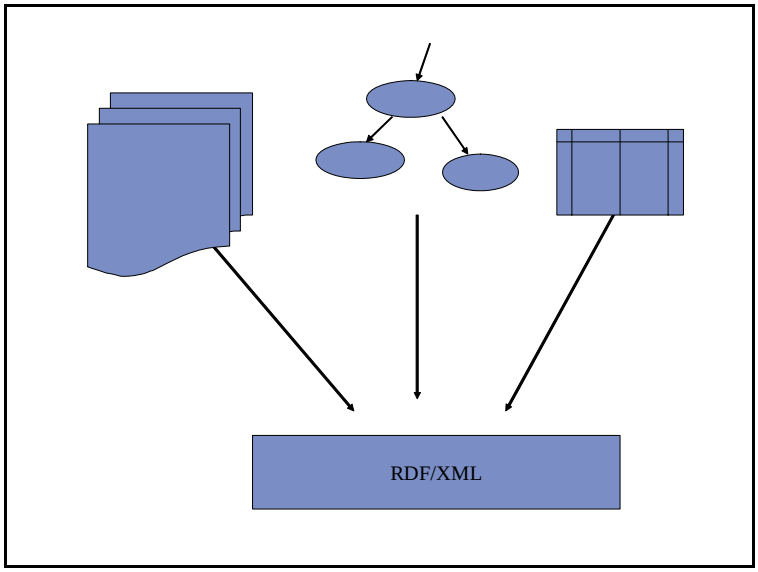

Abbildung 2.1: Heterogenität auf syntaktischer Ebene

Mit RDF in Verbindung mit RDF-S und XML/XML-S als Beschreibungssprache steht ein mächtiges Werkzeug zur Datenmodellierung zur Verfügung.

2.3 Heterogenitätskonflikte zwischen Datenschemata

Auf der strukturellen Ebene stellt sich das Abgleichen unterschiedlicher schematischer Repräsentationen ein und desselben Objekts oder Merkmals als problematisch dar. Dabei unterscheidet man die nachstehenden Konflikte [4]:

- Bilaterlale Konflikte
- Multilaterale Konflikte
- Meta-level Konflikte

Zur Verdeutlichung dieser Konflikte soll im Folgenden ein einfaches Beispiel zur Veranschaulichung dienen. Ein Hotel sei in einer bestimmten Datenquelle durch folgende Repräsentation als RDF-Tripel beschrieben:

```
(http://www.hotels.com#42 name "Amstel Hotel")

(http://www.hotels.com#42 category luxury)

(http://www.hotels.com#42 location "Amsterdam,

 Netherlands")

(http://www.hotels.com#42 priceSingle 250)

(http://www.hotels.com#42 priceDouble 350)
```

Abbildung 2.2: Beispiel 4 [23]

Dabei gilt es zu beachten, dass in anderen Quellen dasselbe Hotel in differierender Form repräsentiert werden kann.

2.3.1 Bilaterale Konflikte

Bilaterale Konflikte betreffen in der Regel genau ein Objekt. Dieses eine Objekt wird in unterschiedlichen Informationsquellen durch verschiedene beschreibende Strukturen abgelegt. Man kann dabei zwischen diesen Konflikten unterscheiden:

Bilaterale Konflikte

- Namenskonflikte (Bezeichnerkonflikte),
- Datentypkonflikte
- Integritätskonflikte

Namenskonflikte

Namenskonflikte treten in allen Fällen auf, in denen Quellen unterschiedliche Namen für dasselbe Objekt der realen Welt (*real world object*) verwenden. Ein typischer Fall ist die Verwendung unterschiedlicher Namen für das gleiche Attribut in relationalen Datenbanken (Synonyme). Ein solcher Konflikt ist in unserem oben dargestellten RDF-Beispiel möglich, so könnten etwa andere Quellen andere Bezeichnungen für die Kategorie (`category`, siehe Abb. 2.2, Zeile 2) eines Hotels z. B. „Klasse", „Sterne" o. ä. verwenden. Ebenso können Homonyme, also syntaktisch gleiche Bezeichner, aber semantisch unterschiedliche Relationen zu Namenskonflikten führen.

Datenkonflikte

RDF unterstützt die Benutzung von XML-Schema-Datentypen, um die einem Objekt zugewiesenen Werte zu repräsentieren. Als **Datentypkonflikt** bezeichnet man den Fall, dass verschiedene Datentypen für denselben Wert verwendet werden, z. B. der Preis für eine Unterkunft einmal im integer-, in einer anderen Datenquelle jedoch im real- oder string-Format angegeben wird. Dies erschwert etwa Vergleichsoperationen.

Integritätskonflikte

Hinsichtlich RDF sind diese Strukturen entweder Ressourcen, Eigenschaften oder Datentypen. Ressourcen werden in RDF durch eine eindeutige Bezeichnung z. B. eine URI gekennzeichnet, in unserem Beispiel `http://www.hotels.com#42.` Dasselbe Objekt kann in einer anderen Repräsentation jedoch durchaus eine andere Bezeichnung tragen, z. B. `http://www.vacation.org/hotels#666.` Die Verwendung unterschiedlicher Identifikatoren für dasselbe Objekt erschwert es, Information über das Objekt aus verschieden Quellen zusammenzufassen und führt zum ***Integritätskonflikt***.

2.3.2 Multilaterale Konflikte

Multilaterale Konflikte

Multilaterale Konflikte sind Konflikte, die mehr als ein Objekt einer Repräsentation berühren. Sie treten auf, wenn eine Information, welche in einer Quelle durch ein Objekt repräsentiert wird, in einer anderen auf mehrere Objekte verteilt ist. Dabei lassen sich drei Konflikttypen unterscheiden:

- Multilateral attribute correspondences
- Multilateral entity correspondances
- Missing values

Multilateral attribute correspondences

Als **multilateral attribute correspondences** bezeichnet man Konflikte, die durch Verteilung von Informationen auf mehrere Merkmale auftreten. In dem

Beispiel ist die Information, wo sich die Unterkunft befindet, über das Merkmal `location` beschrieben.

In anderen Quellen könnte dieselbe Information über zwei Merkmale `city` und `country` gegeben sein. Wird eine Unterkunft an einem bestimmten Ort gesucht, muss die entsprechende Information der verschiedenen Quellen entweder aufgespalten oder kombiniert werden, um sie vergleichbar zu machen (*matchen*).

Multilateral entity correspondances

Als **multilateral entity correspondances bezeichnet** man Konflikte, wenn einzelne oder mehrere Ressourcen verwendet werden, um einen bestimmten Teil der Information darzustellen. In dem vorliegenden Beispiel ist die Information über die Unterkunft und deren Lage in der Beschreibung einer Ressource zusammengefasst. Der Attributwert bzgl. der Lage des Hotels ist hier als Literalwert angegeben. In anderen Quellen könnten spezielle Ressourcen als einmalige Repräsentationen der Lage wie zum Beispiel `cities` und `countries` dienen:

```
(http://www.locations.com/cities\#amsterdam lies_in
http://www.locations.com/countries\ #netherlands)

(http://www.hotels.com#42 location
http://www.locations.com/cities\#amsterdam)
```

Abbildung 2.3 Multilateraler Konflikt

Dieser Zusammenhang wird in **Abbildung 2.**4 noch einmal graphisch veranschaulicht.

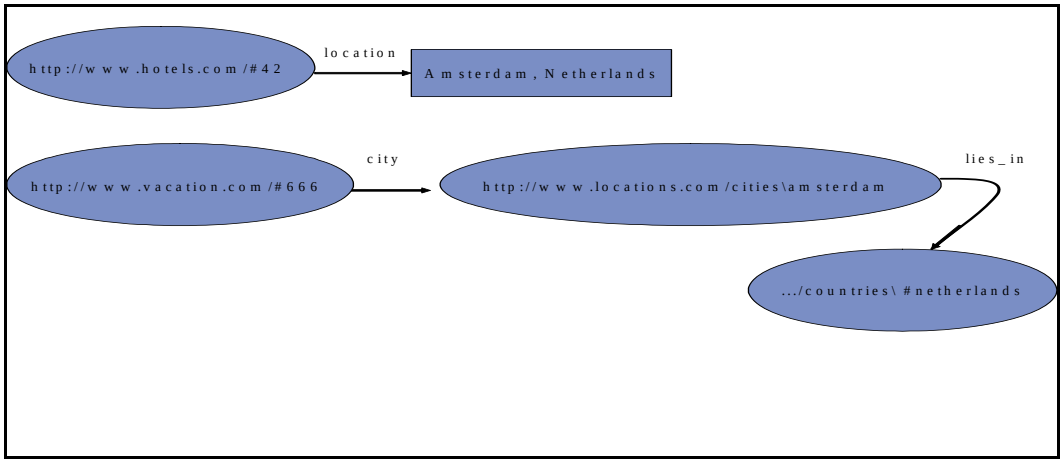

Abbildung 2.4: Multilateraler Konflikt

Missing values

Als **missing values** bezeichnet man Konflikte, wenn bestimmte Angaben in einer Informationsquelle fehlen. In dem oben angeführten Beispiel gibt es Informationen über Einzel- und Doppelzimmer, in anderen Quellen u. U. nur über Doppelzimmer. Dies kann entweder bedeuten, dass es keine Einzelzimmer gibt, dass der Preis identisch ist oder dass diese spezielle Information der Repräsentation nicht eingelesen worden ist.

2.3.3 Meta-Level Konflikte

Meta-Level Konflikte

Diese Konflikttypen sind bedingt durch die Verwendung unterschiedlicher Modellierungselemente zur Repräsentation von Information derselben Art. In konzeptionellen Datenmodellen sind diese Grundelemente Dateneinheiten (*entities*), Attribute und Daten, in RDF-Ressourcen sind es Merkmale und Literale. Die Vermischung bei der Repräsentation kann zu folgenden Konflikten führen:

- Im vorigen Absatz ist deutlich geworden, dass man `location` sowohl als Ressource als auch als Literal repräsentieren kann.

- Außerdem kann eine Situation vorkommen, in welcher dieselbe Information, die in einer Quelle explizit als Relation codiert ist (`netherlands part of europe`), in einer anderen implizit im Typ der Ressource angegeben ist (`netherlands type european country`).

- Darüber hinaus können Datenquellen dieselbe Information beinhalten, aber unterschiedliche Datenstrukturen verwenden. XML-Dokumente können z. B. Elemente in verschiedenen Konstellationen zusammenstellen [1] oder es werden unterschiedliche Wege beschritten, um ein und dieselbe Information darzustellen etwa die Altersangabe in „Alter“ vs. „Geburtsdatum“) [5].

2.4 Heterogenitätskonflikte zwischen Daten-Instanzen

Auf semantischer Ebene stellt sich das Entscheidungsproblem, ob verschiedene Objekte aus unterschiedlichen Quellen dasselbe *real world object* repräsentieren. Ebenso ist danach zu fragen, ob eine gemeinsame Quelle ausgewählt wird, falls widersprüchliche Informationen etwa durch einfache (Tipp-) Fehler erzeugt gefunden werden [1]. Es muss also deren Semantik betrachtet werden, da die beabsichtigten Interpretationen der Repräsentationen von Quelle zu Quelle differieren können. Als Konflikttypen kann man daher auf semantischer Ebene Konflikte unterscheiden, die aus unterschiedlicher Kodierung (**Datenkonflikte**) hervorgehen und solchen, die aus unterschiedlicher Konzeptionierung des jeweiligen Wissensbereichs (**Domänenkonflikte**) erwachsen [4].

2.4.1 Datenkonflikte

Datenkonflikte

Eine Domäne sollte möglichst kompakt und vergleichbar repräsentiert werden. Hierzu werden die wichtigen Eigenschaften der relevanten Objekte beschrieben, indem den einzelnen Merkmalstypen entsprechende Werte zugeordnet werden, die häufig vom konkreten Wert abstrahieren. Je nach Abstraktionsgrad ergeben sich jedoch beim Vergleich verschiedener Repräsentationen u. U. unterschiedliche Auswahlmöglichkeiten entsprechend dieses Werte-Systems, da jede Informationsquelle einen eigenen Maßstab des jeweiligen Werts oder einen eigenen Wertebereich verwendet.

Im Einzelnen spricht man von unterschiedlichen Skalen (*different scales*), wenn insbesondere numerische Werte auf verschiedenen Maßstäben basieren. Ein Beispiel dafür könnte die Angabe des Preises einer Unterkunft je nach Repräsentation in unterschiedlichen Währungen sein. Hinzu kommt, dass die Beziehung zwischen den unterschiedlichen Maßstäben eines Werts (den Währungen) nicht unbedingt konstant sein muss (fester Umrechnungskurs €↔DM vs. variabler Umrechnungskurs €↔$), sodass sie ggf. zum Vergleichszeitpunkt ermittelt werden muss (vgl. Abbildung 2.3).

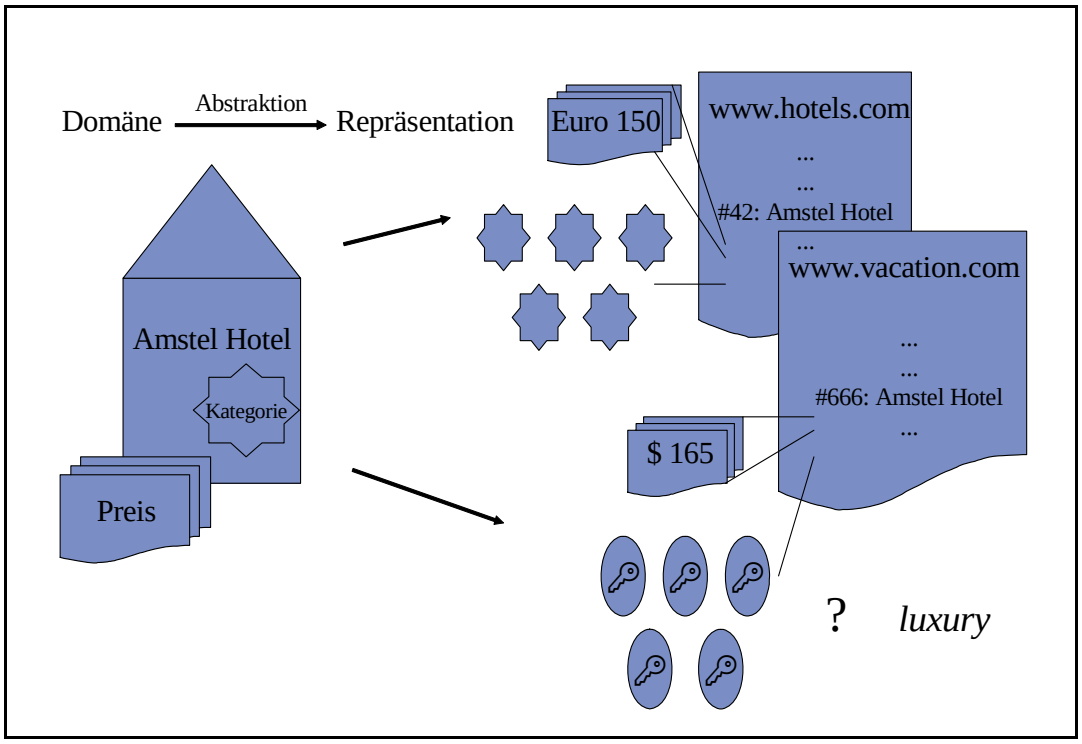

Abbildung 2.3: Datenkonflikte wg. unterschiedlichen Maßstabs (Preise) oder Skalierung (Kategorie); Sterne vs. „`luxury`“ (*surjective mapping*)

Different value ranges

Eine weitere Problematik liegt in der abhängig von der Quelle differierenden Skalierung der Werte. Betrachtet man dies in Bezug auf das Beispiel aus Abb. 2.2, so ergibt sich daraus folgender Konflikt: Die Qualität eines Hotels in Mitteleuropa wird auf einer Skala von einem bis zu fünf Sternen gemessen, in Spanien hingegen verwendet man die Anzahl von Schlüsseln als Qualitätsangabe. Dieser Fall wird als ***different value ranges*** bezeichnet. Besitzt man keine Kenntnis über die zugrunde liegende Skala, ist auch kein Vergleich der Merkmale bzw. deren Abstraktion als Werte möglich.

Surjective mapping

Wenn ein Wert einer Quelle auf mehrere Werte der anderen Quelle abgebildet wird spricht man von dem Konflikttypen des ***surjective mappings***. In Bezug auf das gewählte Beispiel kann dies etwa in der Kategorie `luxury` vier bis fünf Sternen entsprechen. Dabei liegt das Hauptproblem in der Interpretation des Werts. Normalerweise ist es nicht möglich zu entscheiden, ob beispielsweise das Merkmal `luxury` als vier oder fünf Sterne angesehen werden soll. Werden unterschiedliche Repräsentationen für dieselbe Information verwendet, kann die Konvertierung mithilfe von header-Informationen (in XML-Dokumenten) durchgeführt werden. Ein Beispiel dazu könnte die Verwendung sowohl unterschiedlicher Zeichenketten in XML-Dokumenten wie „1“ versus „sehr gut“ als auch unterschiedlicher Zeichensätze sein.

Schema mapping

Schema matching

Die Identifizierung von Objekten bzw. Datenobjekten über Dokumentgrenzen hinweg kann auf Grundlage von ID-Attributen via Referenzierung stattfinden (vgl. XLink, XPointer, [1]). Die Anwendungen, welche die entsprechenden Dokumente erstellen, müssen jedoch sicherstellen, dass diese Mechanismen richtig verwendet werden. Beim Vergleich voneinander unabhängiger Dokumente müssen Elemente, die entweder in Beziehung zueinander stehen oder dasselbe beschreiben, schon während der Integration bestimmt werden. Dieser Konfliktfall wird als ***schema mapping*** oder ***schema matching*** bezeichnet.

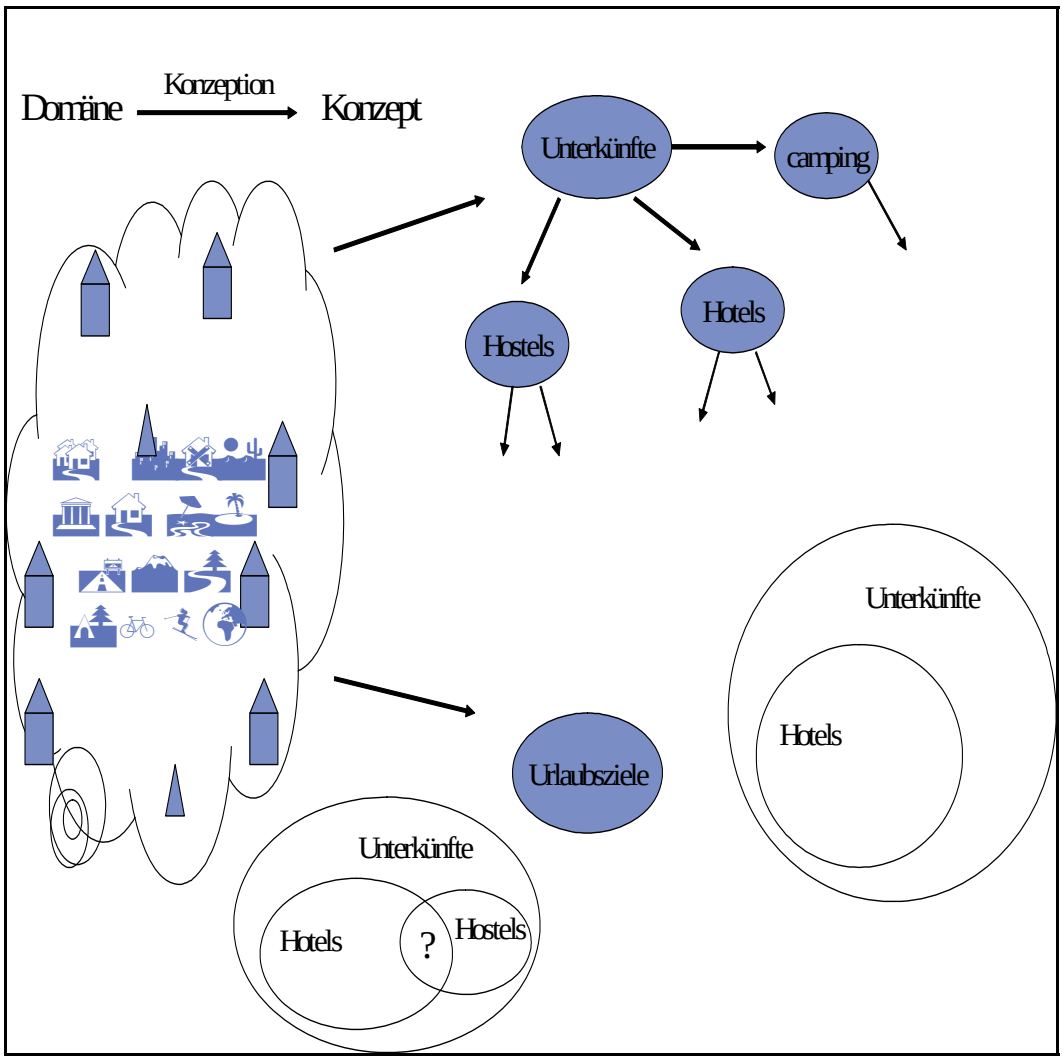

Abbildung 2.4: Domänenkonflikte*: subsumption* (oben), *overlap* (links)

2.4.2 Domänenkonflikte

Domänenkonflikte

Domänenkonflikte treten auf, wenn sich verschiedene Konzeptualisierungen eines bestimmten Objektes der realen Welt widersprechen und somit ein Vergleich unmöglich wird. Diese Art von Konflikten tritt besonders zutage, wenn Objekte und Klassen durch Kategorien voneinander abgetrennt werden, obwohl die Grenzen zwischen den entstehenden Kategorien in der Realität so nicht existieren [6]. Normalerweise werden Objekte in Klassen zusammengefasst, wenn diesen bestimmte Eigenschaften gemeinsam sind, die dann nicht mehr explizit auftauchen. Dadurch wird das Auffinden von Relationen zwischen diesen Kategorisierungen erschwert. Es kann so zu unerwünschter Zusammenfassung, Überlappung und Inkonsistenz von Klassen kommen. Als Dömanenkonflikte kann man daher zwei Typen unterscheiden:

Subsumption

Zu einen ist hier der Fall der **Subsumption*n*** zu nennen. Dieser liegt vor, wenn eine Klasse von Objekten alle Objekte, die in einer anderen Klasse enthalten sind, einbezieht. Für das angeführte Beispiel bedeutet dies, dass die Klasse `accomodation` sowohl alle `accomodation`-Objekte als auch die Klasse `hotels` umfasst. Zwar sind alle Hotels auch Unterkünfte und bei einer Suche nach Unterkünften sollen auch alle Hotels gefunden werden. Eine entsprechende Anfragesprache muss jedoch in der Lage sein, das Datenmodell und die semantischen Eigenschaften des Vokabulars, im Beispiel das Konzept von Unterklassen, zu „verstehen".

Überlappung

Der zweite Konflikttyp stellt den etwas komplexeren Fall der Überlappung vor. Wenn sich zwei Klassen teilweise überlappen (*overlap*), können in dem vorgestellten Beispiel beide Klassen `hotel` und `hostel` einige *hostels* als billige Hotels angesehen werden, während einige Hotels keinesfalls *hostels* sind und einige *hostels* kaum als Hotel gelten können (vgl. Abbildung 2.4). Hier werden zusätzliche Entscheidungskriterien benötigt, welche Teile der Instanzen die Konzepte teilen und welche nicht.

Umgekehrt ist es für den sinnvollen Informationsaustausch nicht nur wichtig zu wissen, wenn zwei Klassen Mitglieder gemeinsam haben, sondern auch, wenn dies gerade nicht der Fall ist, d. h. Klassen per definitionem disjunkt sind, etwa `hotel` und `camp site` in dem vorgestellten Beispiel.

Aggregation

Ein möglicher Konflikt auf Ebene der Domäne (*domain-level*) ist bedingt durch die unterschiedlichen Abstraktionsebenen, die dazu führen können, dass Daten in verbundener Form auftreten (*aggregation*). In dem Beispiel kann etwa eine Quelle Städte entsprechend des Landes, andere entsprechend des Kontinents gruppieren. Häufig ähnelt diese Situation dem Subsumptions-Fall, da z. B. die Klasse `„holländische Stadt“` vom Konzept `„europäische Stadt“` subsumiert wird. Im folgenden Abschnitt werden verschiedene Lösungsansätze vorgestellt, die die hier vorgestellten Konflikte über die Integration auf semantischer Ebene zu beheben versuchen.

2.5 Lösungsansätze für die Semantische Integration

Konkrete Integrationslösungen müssen Konzepte beinhalten, die Konflikte auf allen Ebenen, auf denen sie auftreten können, überwinden [7].

Der Integration auf semantischer Ebene kommt dabei eine besondere Bedeutung zu, da zum einen auf struktureller und syntaktischer Ebene Integrationsansätze und –lösungen Verwendung finden, zum anderen im Web vorrangig unstrukturierte Informationen vorliegen, die einen semantischen Ansatz erfordern.

Die semantische Integration unterstützt das Informationsmanagement, indem sie Daten aus heterogenen, verteilten Quellen beschreibbar und deren Abhängigkeiten untereinander visualisierbar macht. Damit erfolgt eine inhaltliche Integration. Daten aus unterschiedlichen Quellen werden vergleichbar und können ggf. in neue Schemata integriert werden, dadurch wird die Auswahl von Quellen hinsichtlich bestimmter Anforderungen ermöglicht.

Datenintegration

Funktionsintegration

Informationsintegration kann dabei als Verbindung von Daten- und Funktionsintegration angesehen werden. Datenintegration zielt auf die Zusammenführung heterogener Datenbestände ab. Funktionsintegration bezeichnet das Verfügbarmachen lokaler Funktionen bzw. Dienste aus den einzelnen Systemen in einer einheitlichen Form [27]. Das Ziel von Datenintegration ist es, die Daten aus den verteilten, heterogenen Datenquellen zu einer einheitlichen Beschreibung, dem globalen Schema, zusammenzuführen. Dabei handelt es sich um eine vereinheitlichte Sicht der zugrunde liegenden Datenquellen, welche den Benutzern die Möglichkeit bietet, über eine einheitliche Anfrageschnittstelle Anfragen zu stellen, ohne die Heterogenität der Datenquellen und deren Ort berücksichtigen zu müssen [7].

Um Heterogenitätskonflikte aufzulösen, müssen zunächst Übereinstimmungen (sog. m*atches)* zwischen den Schemata der verschiedenen Datenquellen aufgezeigt werden. Dies können 1:1 oder 1:n (bzw. n:n) *matches* sein (vgl. *Schema matching* bzw. *mapping* Kap. 2.4.1). Diese Abbildungsart wird derzeit noch häufig manuell mit Unterstützung verschiedener Anwendungen durchgeführt. Dieser Prozess ist sehr arbeitsintensiv und fehleranfällig. Im Allgemeinen werden 60-80 % der Ressourcen eines Data-Sharing-Projekts für die Auflösung der semantischen Heterogenität verwendet [8].

Zur Verwendung kommen Datenintegrationssysteme, welche im Allgemeinen auf einer wrapper/mediator-Architektur beruhen. Basis ist ein einheitliches und semantisch ausdrucksstarkes Datenmodell zur Repräsentation der Datenquellen. Je nach konkreter Vorgehensweise spricht man vom Materialisierungsansatz, wenn Daten periodisch aus den Quellen extrahiert und in einer zentralen Datenbank abgelegt werden (*data warehouse*). So können Anfragen über diese zentrale Datenbank beantwortet werden. Andernfalls spricht man von dem virtuellen Ansatz [1], der im folgenden Abschnitt näher erläutert wird.

Die Herangehensweise zur Erfassung der Semantik der Daten unterscheidet sich in den verschiedenen Ansätzen (vgl. Kap. 2.5.1.). Der Zugang zur Semantik kann über die Struktur, über die Metadaten der Datenquellen erfolgen oder aber über die natürlich-sprachliche Verarbeitung des Ursprungstextes. Letzteres findet man eher bei unstrukturierten Daten-

quellen, wo entsprechende Integrationssysteme unstrukturierte Anfragen unterstützen müssen (Volltextsuche). Ersteres findet man bei semistrukturierten bzw. strukturierten Quellen, wo die Datenintegration strukturierte Anfragen unterstützt wie in Datenbank-Systemen.

2.5.1 Wrapper und Mediatoren

Wrapper

Wrapper sind Softwarekomponenten, die den Inhalt einer Datenquelle zur Vereinheitlichung in einem anderen Datenmodell oder Schema repräsentieren. Ein Beispiel dafür wäre ein XML-Wrapper für eine relationale Datenbank.

Mediatoren

Mediatoren sind Softwarekomponenten, die der Vereinfachung, Reduzierung, Kombination und Erklärung von Daten dienen. Sie werden v. a. zur Bereitstellung einer gemeinsamen Anfragemöglichkeit auf unterschiedliche Datenquellen genutzt. Abbildung 2.5 stellt diesen Sachverhalt graphisch dar.

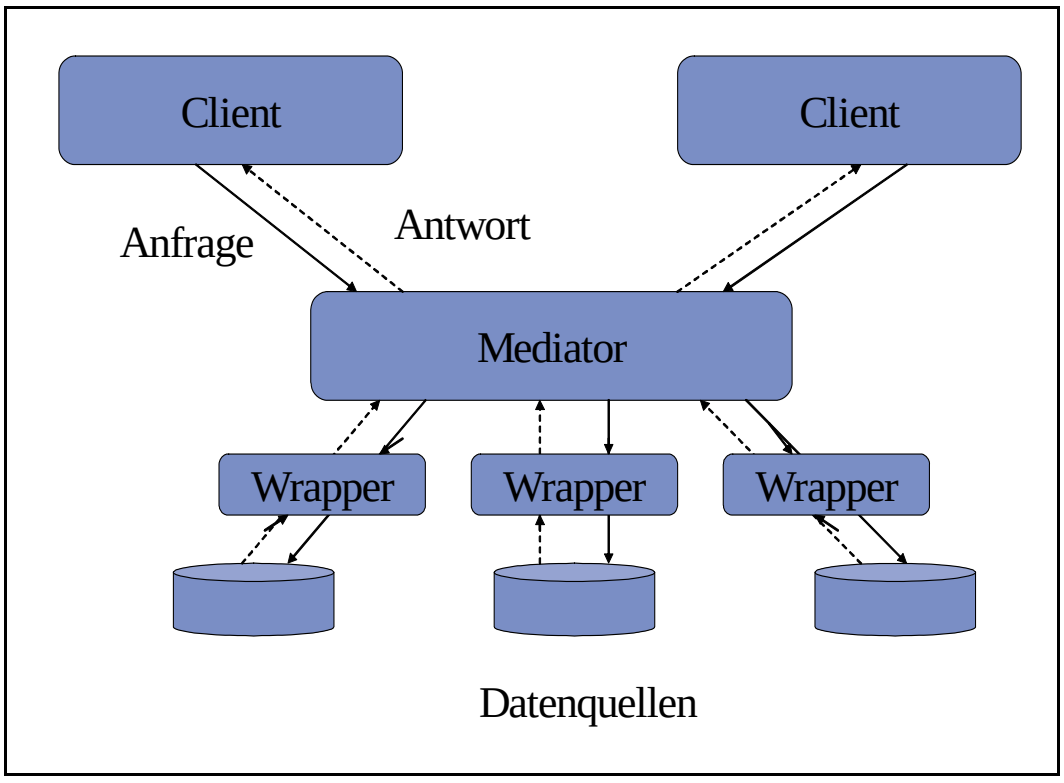

Abbildung 2.5: Wrapper und Mediatoren

Aufgabe des Mediators ist es, Anfragen an das globale Schema in Anfragen an die Quellen zu transformieren sowie die Ergebnisse zu sammeln und zu verknüpfen. Das globale Schema basiert auf einem geeigneten Datenmodell, welches z. B. als XML/RDF Repräsentation vorliegt [9]. Da es sich um eine

virtuelle Integration handelt, werden keine Nutzdaten separat gespeichert, sodass Anfragen an den Mediator unmittelbar in solche an die Quellsysteme umgewandelt werden. Dies geschieht durch Umwandlung der Anfrage in eine dem globalen Schema entsprechende interne Repräsentation und Umschreibung bzw. Zerlegung in durch die Quellsysteme ausführbare Teilanfragen. Die Ergebnisse müssen dann umgekehrt entsprechend der Definition des globalen Schemas und der Anfrage kombiniert und ggf. selektiert oder anderen Operationen unterworfen werden, um anschließend dem Nutzer zur Verfügung gestellt zu werden [10].

Die Kopplung zwischen Quelle und Mediator erfolgt über Wrapper, die dem Mediator einen einheitlichen Zugriff auf die Quellen ermöglichen, indem sie eine Abbildung zwischen dem Datenmodell des Mediators und dem der Quellen erstellen und zudem die Anfragen des Mediators in Anfragen oder Funktionsaufrufe der Quellsysteme übersetzen [11].

Der Zugriff auf Ressourcen strukturierter Information von Wrappern, die aus dem konzeptionellen Modell abgeleitet werden können, kann gewährleistet werden, da in relationalen oder XML-basierten Systemen Standardschnittstellen verfügbar sind. Dagegen stellen wenig strukturierte Informationsquellen wie zum Beispiel HTML-Seiten im Web, ein größeres Problem dar. Die relevanten Daten müssen zunächst identifiziert und extrahiert werden [1]. Hierzu können formale syntaktische Methoden oder Methoden automatischen Lernens verwendet werden. Im zweiten Fall werden repräsentative Beispielseiten mit den relevanten Dateneinträgen dem entsprechenden Lernverfahren vorgelegt. Als Ergebnis des automatischen Lernprozesses gewinnt man eine Menge von Auswahl-Regeln, die zur Extraktion von Informationen aus Web-Ressourcen verwendet werden kann. Diese werden in eine neu geschaffene Struktur eingefügt, die für den weiteren Prozess als Grundlage dient [4].

Damit liegt zwar eine Lösung für das Auslesen von Information aus schwach strukturierten Ressourcen vor, aber das Problem der Integration von Informationen aus verschiedenen Quellen bleibt weitgehend ungelöst, da Extraktionsregeln nur auf struktureller Ebene definiert sind. Man benötigt also eine Verknüpfung zwischen zwei Ebenen, zum einen der Ebene der extrahierten Information und zum anderen der Ebene des jeweiligen Datenmodells.

2.5.2 Erfassen von Semantik über die Struktur

Ein weit verbreiteter Weg, um die Bedeutung von Information zu erfassen, ist die Beschreibung ihrer Struktur in Ausdrücken. Die Verwendung von konzeptionellen Modellen der gespeicherten Information ist von Datenbanksystemen bekannt (*Entity-Relationship-Modell*). Hier werden Objekte und Beziehungen der abzubildenden Domäne speziellen Tabellen, genauer gesagt

deren Spalten oder Zeilen, zugeordnet (siehe Abbildung 2.6). Solche konzeptionellen Modelle weisen eine enge Verbindung zu der Art der Speicherung von Informationen auf. Sie strukturieren also die abzubildende Domäne ohne Berücksichtigung von situationsbezogenen und semantischen Unschärfen bei der Modellierung von Realweltausschnitten. Dadurch erfolgt eine Reduktion auf präzise, vollständig definierte Daten und Attribute, dies kann zu Informationsverlust bzw. -veränderung führen. Die Verknüpfung von Struktur und semantischem Modell hat jedoch bedeutende Vorteile für die gemeinsame Nutzung von Informationen. Das konzeptionelle Modell unterstützt bei dem Zugriff auf Information und bei deren Validierung [4]. Für die Integration auf semantischer Ebene muss auf der strukturellen Ebene aufsetzend ein logisches (Daten-) Modell erstellt werden, d. h. das Datenmodell wird in einen bestimmten Kontext gestellt. Dabei können entsprechende Modelle ganzen Datenbank-Schemata entsprechen oder auch nur einzelnen Termen, die in der Datenbank verwendet werden. Aufgabe der Informationsintegration ist, wie oben beschrieben, die Herstellung eines einheitlichen virtuellen oder materiellen Schemas, damit Quellen trotz der ihnen zugrunde liegenden unterschiedlichen konzeptionellen Modelle vergleichbar werden und Anwendungen auf ihnen wie auf einer einzigen, integrierten Informationsquelle zugreifen können.

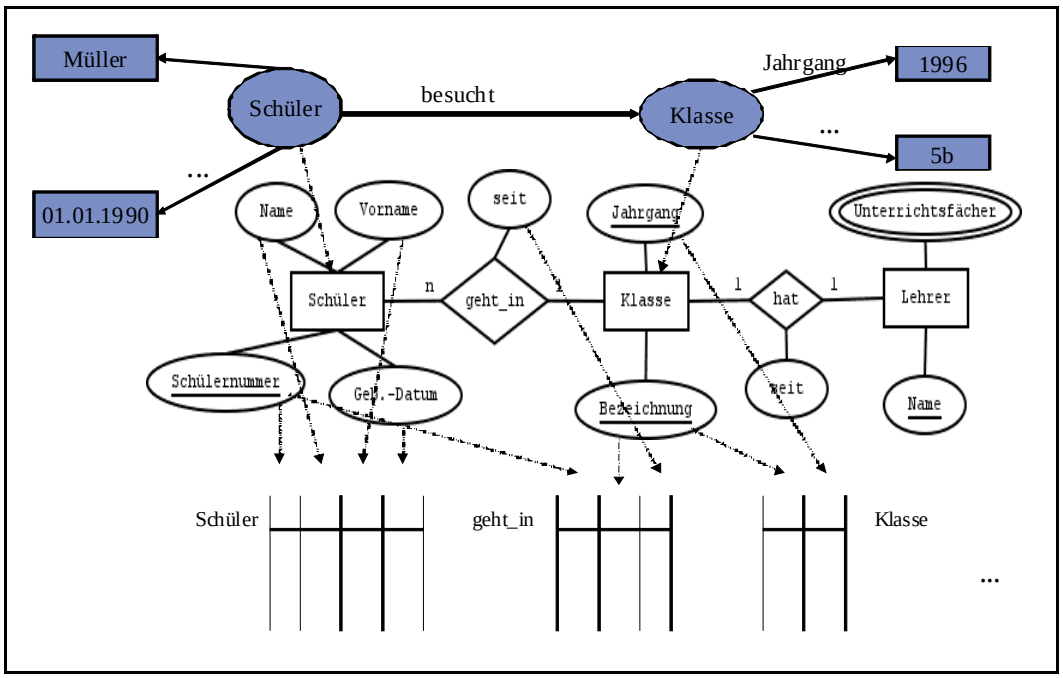

Abbildung 2.6: Erfassen der Bedeutung über die Struktur (ERM aus www.tinohempel.de/info/info/datenbank/erm.htm; letzter Zugriff 2007/02/02)

Hierzu wird es nötig sein, die unterschiedlichen Konzepte miteinander zu homogenisieren (*matchen*). Dies ist semantisch korrekt nur dann möglich, wenn allen zu integrierenden Datenquellen dasselbe semantische Modell zugrunde liegt, sodass alle Elemente und Beziehungen des einen Konzepts ihre Entsprechungen im anderen finden. Abbildung 2.7 soll diesen Sachverhalt veranschaulichen.

Es gibt unterschiedliche Ansätze zur Verknüpfung von Informationsquellen und Datenmodellen [12], die im Folgenden dargestellt werden.

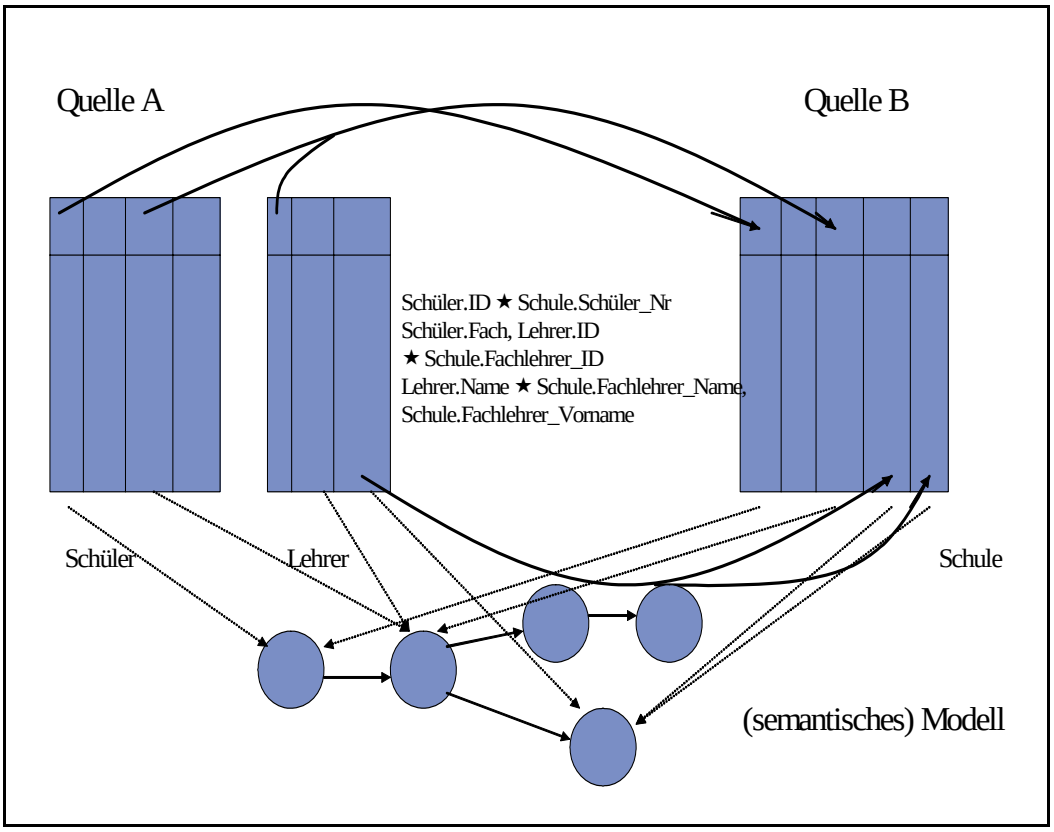

Abbildung 2.7: Matching über gemeinsames semantisches Modell

2.5.2.1 Struktur-Ähnlichkeit (*structure resemblance*)

Es wird ein logisches Modell erstellt und in eine Sprache kodiert, die automatisches, logisches Schließen ermöglicht. Das Modell stellt eine 1:1 Kopie der konzeptionellen Struktur der Datenbank dar. Die Integration erfolgt auf der Kopie dieses Modells und kann anschließend leicht auf die Originaldaten zurückgeführt werden. Diese Methode ist etwa im SIMS-Mediator, dem MOMIS System [13] und im TSIMMIS System [14] implementiert. Eine geeignete Kodierung der Informationsstruktur kann bereits genutzt werden, um Hypothesen über die semantisch zusammenhängenden Strukturen in zwei Informationsquellen zu erstellen [4].

2.5.2.2 Term-Definition (*definition of terms*)

Um die Semantik von Ausdrücken in einem Datenbankschema greifbar zu machen, reicht eine Kopie desselben nicht aus. BUSTER [15] stellt beispielsweise einen Ansatz dar, der ein Datenmodell zur Definition von Termen aus der Datenbank oder dem Datenbankschema verwendet, die nicht unmittelbar mit deren Struktur korrespondieren, sondern lediglich mit den entsprechenden Informationen verlinkt sind. Die Definition selbst kann sogar aus einer Reihe von Regeln bestehen, die den Term definieren. In den meisten Fällen wird er jedoch von Konzeptdefinitionen beschrieben [11].

2.5.2.3 Struktur-Anreicherung (*structure enrichment*)

Hier wird ein logisches Modell konstruiert, welches die Struktur der Informationsquelle abbildet und zusätzlich Definitionen von Konzepten enthält. Es stellt momentan den verbreitetsten Ansatz dar [16] und verbindet die beiden oben genannten Ansätze. Zu den Systemen, die diese Methode nutzen, gehören etwa OBSERVER, KRAFT, PICSEL, DWQ und InfoQuilt [4, 17]. OBSERVER verwendet Ontologien (vgl. Kurseinheit 2), um Informationsquellen zu beschreiben und Beziehungen (Synonyme, Hypo-, Hypernyme) zwischen Termen unterschiedlicher Modelle zu ermöglichen zur Anfragenübersetzung. Dieser Ansatz ist beschränkt auf einfache Beziehungen [4]. PICSEL und DWQ definieren die Struktur der Informationen mit typisierten Horn-Regeln. Die darin enthaltenen zusätzlichen Konzept-Definitionen werden von einem Beschreibungs-Logik-Modell (*description-logic model*) erzeugt. KRAFT legt kein spezielles Definitionsschema fest [15]. InfoQuilt ermöglicht die Darstellung von komplexen Beziehungen zwischen verschiedenen Domänen, unterstützt verschiedene Funktionen und Simulationen und stellt so ein mächtiges Abfrage-Interface zur Verfügung [18].

2.5.2.4 Meta-Annotation

Dies ist ein relativ junger Ansatz, der den Gegebenheiten des Webs Rechnung trägt, wo semantische Informationen häufig in Form von Kommentaren bzw. Anmerkungen hinzugefügt werden. Diese Konstrukte werden genutzt, um Zugang zur Semantik zu gewinnen. Ontobroker und SHOE sind Beispiele für diesen Ansatz [12].

Da Semantik nur bedingt aus der Struktur ableitbar ist, weil in der Regel kein konzeptionelles Modell zugrunde liegt, ist der Ansatz, sich der Semantik über die Struktur der Datenquellen zu nähern, für das Web oft ungeeignet (vgl. [5]). Das Web stellt ja gerade ein sehr umfangreiches Reservoir an nicht oder wenig strukturierten Inhalten dar. Wie schon erwähnt bietet sich hier der folgende Zugang an.

2.5.3 Zugang über natürlich-sprachliche Verarbeitung des Ursprungstextes

Information Retrieval

Hierbei finden Methoden des Information Retrieval Anwendung. Die Aufgabe, relevante Informationen zu einem bestimmten Thema durch inhaltliche Suche auf Dokumenten zu finden, wird durch Indexierung von freien Text-Dokumenten mit gewichteten Ausdrücken, die mit ihrem Inhalt verbunden sind, gelöst. Die etwa von Suchmaschinen im Web eingesetzten Systeme zur Informationsgewinnung erfassen Eigenschaften und Charakteristika von Dokumenten, ohne diese zu verwalten. Unter dem Begriff des Information Retrieval ist gewöhnlich die vage Suche auf Dokumentinhalte und deren unspezifische Bewertung zu verstehen [vgl. 1].

Deskribierung

Deskribierung der Texte, Recherche und Bewertung der Ergebnisse sind die einzelnen Phasen des Information Retrieval. Die Deskribierung beschreibt die Transformation eines Textdokuments in eine Dokumentbeschreibung aufgrund von Metadaten und Schlagworten über Stichworte aus dem Text, den so genannten indexierten Termen. Die Analyse des Textes anhand der vorkommenden Worte wird auch Indexierung genannt. Statistische, wortbasierte Verfahren sind dabei am gebräuchlichsten [4].

Disambiguierung

Da der semantische Inhalt eines Dokuments in den indexierten Termen beinhaltet ist, stellt deren Wahl und Erstellung bei der Behandlung der Semantik des Textes den entscheidenden Schritt dar. Dokumente werden durch den Abgleich der Terme von deren Dokumentbeschreibungen mit einem semantischen Modell, das kontextuelle Zusammenhänge und Beziehungen der Terme berücksichtigt, vergleichbar. Damit steht ein Verfahren analog zum *Matchen* von Strukturelementen aus dem vorigen Abschnitt zur Verfügung. Die Aufgabe, die beabsichtigte Bedeutung der Terme im jeweiligen Zusammenhang zu ermitteln, wird als Disambiguierung (*disambiguation*) bezeichnet.

Die Auswahl der Terme lediglich anhand einfacher linguistischer Analysen, die Berücksichtigung von Beugungsformen kann hier als Beispiel genannt werden,3 und Bewertung anhand der Häufigkeit eines Worts führt oft zu unbefriedigenden Ergebnissen. Beispielsweise kann ein Wort sowohl als Verb als auch als Adjektiv (*fabricated units vs they fabricated units*) eingestetzt werden, dies kann zu unterschiedlich starkem Zusammenhang mit Benutzeranfragen führen. Daher analysieren komplexere linguistische Verfahren Worttypen und Satzstrukturen derart, um Worte in ihrem Zusammenhang zu erkennen. Hierzu werden bestimmt linguistische Eigenschaften der Wörter wie Subjekt, Präposition etc. berücksichtigt. Eine zusätzliche Verbesserung der Erfassung semantischer Zusammenhänge kann durch Betrachtung des jeweiligen Kontextes, in dem ein Term vorkommt, erreicht werden. Dies erfolgt unter Berücksichtigung und Entscheidung zwischen verschiedenen

Latent semantic

möglichen Interpretationen, basierend auf dem Vorkommen anderer Wörter in diesem Zusammenhang. Daraus kann sich zugleich ein Hinweis auf eine bestimmte Bedeutung ergeben. Die Ausnutzung dieser impliziten Strukturen wird als verborgenes semantisches Indexieren (*latent semantic analysis*) bezeichnet.

Die Entscheidung für eine bestimmte Bedeutung basiert häufig auf einem natürlichsprachlichen Wortschatz etwa in Form eines Lexikons oder eines Thesaurus. Zudem kann ein Thesaurus zu Termen noch verschiedene Beziehungen wie Vorzugsbenennungen, Querverweise, Ober- und Unterbegriffe u.s.w. speichern, um Fälle, in denen spezialisierte Vokabulare in Dokumenten verwendet werden, zu berücksichtigen (siehe Abbildung 2.8).

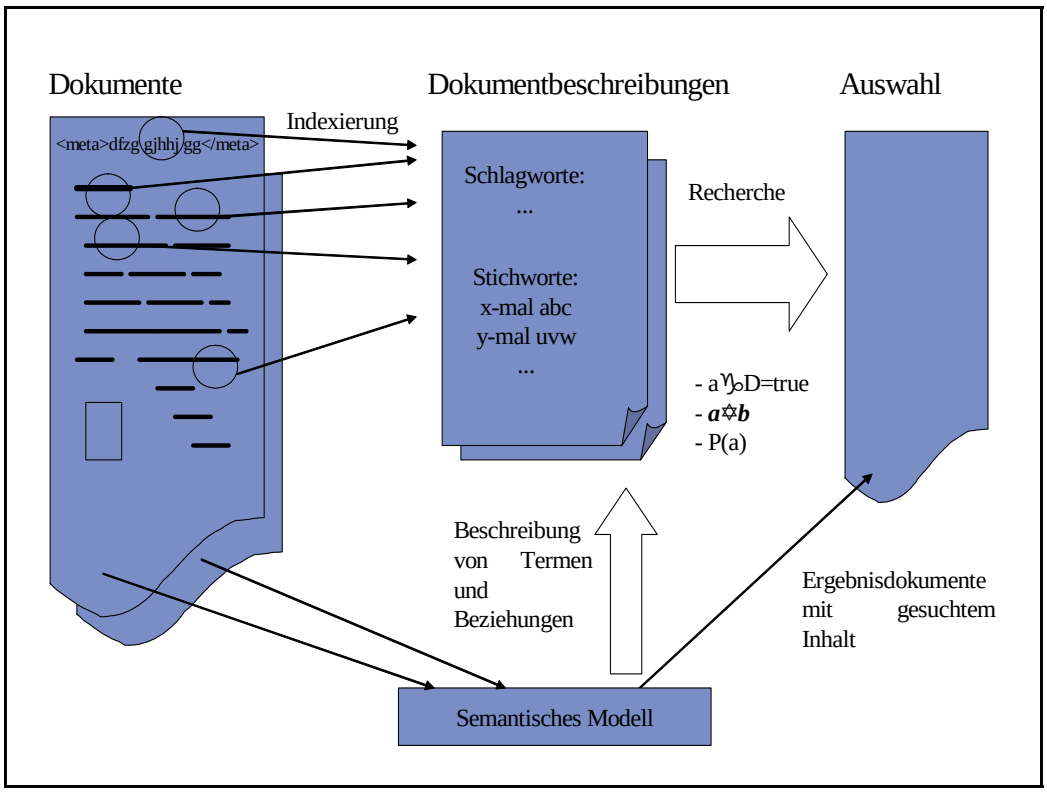

Abbildung 2.8: Erfassen der Bedeutung über natürlichsprachliche Verarbeitung

Diese Beziehungen werden von domänenspezifischen Lexika oder semantischen Netzen zur Verfügung gestellt. Ausdrucksstärkere Terme zur Gewinnung von komplexer indexierter Information erhält man durch Verwendung von Ontologien (vgl. Kurseinheit 2), die das Anwendungsgebiet geeignet strukturieren und so komplexere Zusammenhänge darstellen können.

Nachdem die Dokumente durch eine Deskribierung entsprechend aufbereitet sind, können Anfragen an die Dokumente gestellt werden. Bei der Recherche werden die gewonnenen Terme (Stichworte, Schlagworte) verwendet und mit verschiedensten Operatoren, Kontextinformationen oder Wahrscheinlichkeiten versehen, um die Menge der Ergebnisdokumente effektiv einzuschränken.

Boole´sches Retrieval

Eine Auswertung der Recherchen kann nach verschiedenen Retrieval-Modellen erfolgen. Im einfachsten Fall, dem Boole'schen Retrieval, gibt der Wert `true` an, dass die recherchierten Terme im Dokument vorkommen. Mehrere gefundene Dokumente haben ggf. dieselbe Relevanz.

Vektorraum-Modell

Im Vektorraum-Modell werden die Such- und Dokument-Terme jeweils als Vektoren im mehrdimensionalen Suchraum aufgefasst. Mit vorgegebenen Ähnlichkeitsmaßen können auch Dokumente mit „benachbarten" Vektoren gefunden werden, falls die Such-Terme nicht direkt getroffen werden. Es wird also ein mehr oder weniger vages Ergebnis geliefert, welches durch Wahrscheinlichkeiten der Terme verfeinert werden kann. In probalistischen Modellen werden Wahrscheinlichkeiten hinsichtlich der Relevanz eines Dokuments in Abhängigkeit zur Anfrage berechnet.

Relevance Feedback

Beim Ranking sollen schließlich die gefundenen Dokumente entsprechend ihrer Relevanz sortiert werden, ggf. kann der Nutzer beim *Relevance Feedback* bestimmte Dokumente entweder als irreleant oder als relevant markieren und so die Rankingsortierung hin zu relevanteren Dokumenten verschieben.

Eine Besonderheit des Information Retrieval im Web stellt die Tatsache dar, dass nicht von bestehenden Dokumentmengen ausgegangen werden kann, sondern diese im Web eingesammelt werden müssen (*crawling*). Hierzu werden die Link-Referenzen zwischen den Dokumenten ausgenutzt. Die ermittelte Link-Struktur kann zu Ranking-Zwecken verwendet werden. Generell stellen sich an Retrieval-Techniken im Web spezielle Anforderungen, zu denen nicht nur eine möglichst vollständige Abdeckung der verfügbaren Information, sondern auch die Aktualität der Indizes zählen. Ersteres ist angesichts des Umfangs des Web-Inhalts kaum realisierbar. Tatsächlich werden nur etwa 60 % der direkt erreichbaren Dokumente und nur ein Bruchteil der in Web-Datenbanken vorgehaltenen und ggf. über dynamisch generierte Webseiten ausgegebenen Information gefunden. Die Aktualität der Indizes liegt bei etwa 50-150 Tagen, d. h. jüngere oder in dieser Zeit geänderte Webinhalte werden u. U. nicht gefunden [1].

Die Methoden des Information Retrieval finden ihre Grenzen zum einen bei spezialisierten Texten, insbesondere der wissenschaftlichen Art, da die dort

verwendete Terminologie gar nicht oder kaum mit bestehenden linguistischen Verfahren erfassbar ist und zum anderen bei komplexerer Indexierung. Hier fehlt ein eindeutiges semantisches Modell, welches Terme und ihre Beziehungen beschreibt.

2.5.4 Semantische Modelle

Spezielle Vokabulare treten, wie oben gezeigt, oft in Klassifikations- und Bewertungsausdrücken auf. Sie dienen der Reduzierung des in der Informationsquelle gespeicherten Datenaufkommens. Statt alle Merkmale eines durch einen Datensatz repräsentierten Objekts zu beschreiben, wird eine Klasse verwendet, die Objekte mit denselben Eigenschaften beinhaltet. Dieser Ausdruck entspricht oft einer Klassifizierung, welche außerhalb der Informationsquelle spezifiziert wird, z. B. die Verwendung von Produkt-Kategorien bei e-Commerce oder der Bezug auf standardisierte Flächennutzungsarten in geographischen Informationssystemen [19]. Wie bereits erwähnt, können so wichtige Information verloren gehen.

Klassifizierung

Neben der großen Bedeutung für die gemeinsame Informationsnutzung (*information sharing*) im Web, ist der Einsatz von heterogenen Klassifikationsschemata auch für die Integration von Information bedeutsam. Information Sharing erfordert sowohl das Auffinden geeigneter Informationsquellen als auch den Zugang zu den darin enthaltenen Daten, d. h. die gefundenen Informationsquellen müssen mit dem anfragenden System zusammenarbeiten können. Man spricht hier von der informationellen Interoperabilität [12]. Hierzu sind auf niedrigeren Ebenen formale Sprachen wie XML oder RDF nötig, die sowohl die Struktur der Information als auch Meta-Informationen über die Natur der Information und der konzeptionellen Struktur, die der Informationsquelle zugrunde liegen, erfassen und beschreiben können (vgl. Kurs 01873). In einem weiteren Schritt müssen die Informationsquellen mithilfe der erwähnten Sprachen mit Information zur Semantik versehen werden. Diese semantische Information muss auf einem Vokabular basieren, welches eine übereinstimmende und formale Spezifikation der Konzeptualisierung der Domäne zum Ausdruck bringt [4]. Bei der Integration von Information können diese geeigneten semantischen Modelle verwendet werden, um die Semantik der Informationsquelle zu beschreiben und offen zu legen.

Information sharing

In Hinblick auf die Integration von Datenquellen können sie zur Identifizierung und Verbindung von sich semantisch entsprechenden Informationskonzepten verwendet werden (*semantic matching*). Diese Ebene umfasst insbesondere das Ausstatten der Informationsquellen mit zusätzlicher semantischer Information und der Verwendung von gemeinsam genutzten Term-Definitionen. Dies wird bereits in aktuellen Ansätzen gemeinsamer Informations-Nutzung in Form von Meta-Notationen und Term-Definitionen

implementiert [4]. Neben der Beschreibung und Offenlegung der Semantik von Informationsquellen können semantische Modelle auch als globales Anfragemodell und Grundlage für die Verifikation von Integrationsschritten dienen [12].

2.6 Semantische Modelle darstellen und vergleichen

Wissensrepräsentation

Die Integration von Information wirft die Frage nach der Natur der verwendeten semantischen Modelle auf. In der Regel liegen einzelnen Informationsdarstellungen unterschiedliche Modelle zugrunde. Maßgeblich ist hierbei deren formale Repräsentation, d. h. die Art der verwendeten Sprache. Kern derselben ist ein im letzten Abschnitt vorgestelltes Model der Wissensrepräsentation [12].

Wesentliche Voraussetzung für die Integration von Information ist die Vergleichbarkeit von Information auf semantischer Ebene. Dem liegt die Vergleichbarkeit der Bedeutungen von Ausdrücken zugrunde, die als Elementnamen eines Schemas und als Attributwert verwendet werden. Ein einfaches Beispiel für die Problematik, Ausdrücke in Relation zu einer bestimmten Bedeutung zu setzen, zeigt sich bei der Verwendung von Wörterbüchern. Je nach Kontext kann ein Wort unterschiedliche Bedeutungen haben (sog. Homonyme). Andererseits können unterschiedliche Worte in einem bestimmten Kontext Bedeutungsgleichheit bzw. Bedeutungsähnlichkeit besitzen (sog. Synonyme). Durchsucht man Dokumente nach bestimmten Ausdrücken ohne Berücksichtigung ihrer Bedeutung, kann man demzufolge sowohl irrelevante, also zu viele Ergebnisse als auch zu wenige Ergebnisse erhalten. Basis für Integration und Transformation von Datenwerten stellen die semantischen Beziehungen zwischen diesen Ausdrücken dar [4].

Hierbei gibt es unterschiedliche Möglichkeiten der Konzeptualisierung und der Kontext-Wissens-Darstellungen. Die Möglichkeiten reichen mit wachsender semantischer Ausdrucksstärke von der einfachen informellen Beschreibung von Ausdrücken in natürlicher Sprache (Glossar) über einfache Hierarchien von Ausdrücken oder komplexe Netzwerke über Hierarchien von Konzepten, komplexen Konzeptbeschreibungen (Ontologien) bis hin zu streng formalen Ansätzen mit der Ausdrucksstärke von Prädikatenlogik [15].

Die einzelnen Ansätze sollen im Folgenden etwas näher beschrieben werden (siehe Abbildung 2.9). Hierzu wird als Beispiel der Begriff `trip` betrachtet, der dem Wörterbuch nachfolgende Bedeutungen besitzen kann:

1. a journey for some purpose (he took a trip to the shopping center)
2. a hallucinatory experience induced by drugs (an acid trip)
3. an accidental misstep threatening or causing a fall
4. tripper, trip: a catch mechanism/switch (the pressure activates the tripper and releases the water)
5. a light or nimble tread (he heard the trip of women's feet overhead)
6. an unintentional but embarassing blunder (recite a poem without a single trip)

Abbildung 2.9: Beispiel 5 [4]

2.6.1 Namen und Bezeichner

Die im zweiten Abschnitt angesprochene Möglichkeit der Definition von Terminologien mittels DTD bzw. XML-S in XML-Dokumenten stellt letztendlich nichts anderes als eine Hierarchie von Termen dar. Allgemeiner formuliert repräsentiert die Organisation von Termen in Netzwerken durch zweiwertige Relationen die einfachste Art, einen semantischen Kontext auszudrücken. Vor allem auf dem Gebiet des Information Retrieval ist eine Reihe von Methoden zur Kontextsklärung entwickelt worden, die weitere Ausdrücke verwenden, um mehr Information über die beabsichtigte Bedeutung eines Ausdrucks zu liefern. Ein Ansatz dazu kann in der Verwendung von Mengen von Synonymen anstelle einzelner Ausdrücke gesehen werden. Ein Synonyme-Set enthält alle Ausdrücke, denen eine Bedeutungsgleichheit bzw. –ähnlichkeit zueigen sind. In dem oben angeführten Beispiel wären demnach `trip` und `journey` in einem Synonyme-Set mit der ersten Bedeutung, ein Synonym-Set mit der zweiten Bedeutung würde z. B. `hallucination` enthalten.

Semantischer Kontext

Mithilfe von Ähnlichkeitsmaßen, welche allen Mitgliedern der Synonyme-Sets zweier zu vergleichender Ausdrücke zugewiesen werden, können Ausdrücke mit gleicher bzw. ähnlicher Bedeutung gefunden werden, da ihre Synonyme-Sets einige Ausdrücke gemeinsam haben. Sie verhindern außerdem falsche Übereinstimmungen zwischen Ausdrücken unterschiedlicher Bedeutung, da deren Synonyme-Sets annähernd disjunkt wären.

2.6.2 Termnetzwerke

Nutzt man weitere Relationen wie etwa Hyper-, Hypo-, Holo-, Mereonyme zu anderen Termen kann man Netzwerke aufbauen, die den Kontext eines Terms weit ausführlicher beschreiben.

Den Oberbegriff eines Begriffes nennt man Hypernym. Unter Hyponym versteht man den Unterbegriff eines Begriffs. Dabei ist der Begriffsumfang

Hypernym

Hyponym

des Hyponyms kleiner als der des Hyperonyms, aber der Begriffsinhalt des Hyponyms ist größer als der des Hyperonyms.

Holonym

Meronym

Der Begriff der Holonymie bezeichnet das Wort einer „Teil-von-Beziehung" zwischen Wörtern, das das andere mit beinhaltet. Die Umkehrung dieser Relation bezeichnet man als Meronymie. So stellt der Begriff Finger ein Meronym von Hand dar und der Begriff Hand ist ein Holonym von dem Begriff Finger.

Thesaurus

Die verbreitetste Form solcher Netzwerke sind Thesauri (Synonymwörterbücher), die hauptsächlich Hyper- und Hyponym-Relationen verwenden, um Hierarchien von Ausdrücken zu bilden. Auch hier können Ähnlichkeitsmaße helfen, semantische Zusammenhänge aufzudecken. Zur Ermittlung der Ähnlichkeitsmaße kann man unter anderem die Pfadlänge zwischen zwei Ausdrücken sowie statistische Informationen nutzen.

2.6.3 Konzepthierarchien

Das Problem von Term-Netzwerken ist, dass sie kein formales Prinzip zum Hierarchieaufbau anbieten. Daher können immer noch verschiedene, mögliche Interpretationen eines Ausdrucks denselben Platz in einer Hierarchie teilen. In Bezug auf das Beispiel in Abbildung 2.10 stehen `trek` und `tumble` (Sturz wg. `trip` u. a. Stolpern) auf gleicher semantischer Ebene im Hinblick auf `trip` . Nutzt man anstelle einer Hierarchie von Ausdrücken eine Hierarchie von Konzepten zur Beschreibung der Bedeutung der Ausdrücke, kann jedes Konzept durch die Eigenschaften seines Vorgängers in der Hierarchie, die es erbt, definiert werden. Die geerbten Definitionen können beim Vergleich der Bedeutung von zwei Konzepten benutzt werden, was mehr und genauere Informationen liefert.

2.6.4 Ontologien

Fügt man den Instanzen der Konzepte bestimmte Kennzeichen und Einschränkungen hinzu (*features, constraints*), kann man deren Bedeutung noch näher und aussagekräftiger bestimmen. Im oben angeführten Beispiel könnte definiert werden, dass jeder `trip` Attribute wie z. B. eine Richtung (*destination*) und eine Dauer (*duration*) besitzt, dass er aus verschiedenen Teilen bestehen kann entweder im Sinne von Abschnitt/Etappen (*stages*) oder im Sinne von Etappe/Runde (*legs*) und dass er verschiedene Funktionen erfüllen kann wie etwa einen Besuch (*visit*). Ein formaler und verbreiteter Ansatz nutzt RDF-S (Vgl. Abschnitt 2), um Konzepte entsprechend zu beschreiben. Ein RDF-Schema enthält Definitionen von Klassen mit entsprechend zugehörigen Eigenschaften, die durch *constraint-properties* enger gefasst werden. Default-Werte und Einschränkungen des Wertebereichs sind jedoch immer noch von zu geringer Ausdrucksstärke. Die Definition von Klassen mittels logischer Formeln, Regelmengen oder komplexer Axiom-

systeme in Logik erster Ordnung bietet weit mächtigere Repräsentationsmöglichkeiten. Die Entwicklung spezieller Repräsentationsformalismen bieten epistemologische (erkenntnistheoretische) Primitive zur Definition von Konzepten mittels Kennzeichen von deren Instanzen an. Zu den am häufigsten verwendeten gehören frame-basierte Repräsentationen und Beschreibungslogiken (*description logics*). Frame-basierte Systeme definieren einen Rahmen mit fester Struktur zur Beschreibung der Eigenschaften der Instanzen von bestimmten Konzepten, während *description logics* eine Untermenge der Prädikatenlogik darstellen. Sie bieten eine flexible logische Sprache zur Definition notwendiger und hinreichender Bedingungen an, die bestimmte Instanzen erfüllen müssen, um zu einem Konzept zu gehören [18].

Alle erwähnten Ansätze zur Beschreibung von Semantiken, die auf Merkmalen von Instanzen beruhen, können auch zu deren Vergleich eingesetzt werden. Auf dem Gebiet des fallbasierten Schließens (*reasoning*) sind Ähnlichkeitswerte definiert worden, die den Vergleich von Konzepten erlauben, die als auf Attribut-Wert-Paaren basierende „Fälle" repräsentiert werden. Für frame-basierte Sprachen sind Vergleichsalgorithmen vorgeschlagen worden, die durch die Berücksichtigung der Struktur der Konzeptausdrücke semantische Übereinstimmungen ermitteln können. Im Fall der 1. Ordnung-Axiomsysteme (Prädikatenlogik 1.Ordnung) kann logisches Schließen genutzt werden, um festzustellen, ob eine Repräsentation eine andere impliziert oder ob zwei Repräsentationen äquivalent sind und daher dieselbe Bedeutung haben.

Da diese Art von Vergleichen von Semantiken, die auf dem Prinzip allgemeiner Deduktion sich oft als schwierig erweisen, bietet die Beschreibungslogik spezialisierte Schließverfahren, um zu ermitteln, ob die Definition eines Konzepts ein Spezialfall eines anderen ist. Diese Möglichkeit macht Beschreibungslogiken zu einem mächtigen Werkzeug zum Beschreiben und Vergleichen von Semantiken und zum verbreitetsten Ansatz zur Darstellung von Ontologien [20].

2.7 Zusammenfassung

Die Erschließung und Nutzung der umfangreichen Informationsquellen im Web ist nur mit maschineller Unterstützung möglich. Hierzu ist es erforderlich, die vorhandenen Informationen für maschinellen Zugang und Verarbeitbarkeit zu öffnen. Wesentliche Voraussetzung hierfür ist die Integration heterogener, verteilter Informationen und die Erschließung ihrer Semantik. Für eine Erfassung der Semantik der vorliegenden Informationen ist eine geeignete Darstellungsform notwendig. Syntaktische Standards wie XML und RDF und die Verwendung von strukturierten Metadaten zur „Anreicherung" der entsprechenden Dokumente mit semantischem Inhalt ermöglichen in

Verbindung mit semantischen Modellen und den entsprechenden (Ontologie-) Sprachen den Zugang entsprechender Software-Tools zur Semantik der Web-Inhalte.

Die Entwicklung erfolgt auf voneinander unabhängigen, jedoch aufeinander aufbauenden Ebenen gemäß eines geeigneten Schichtenmodells, wie es beispielsweise vom W3C entwickelt worden ist. Die vorliegende Kurseinheit hat zunächst die untere Hälfte des Schichtenmodells (vgl. Abb. 10) im Überblick beschrieben.

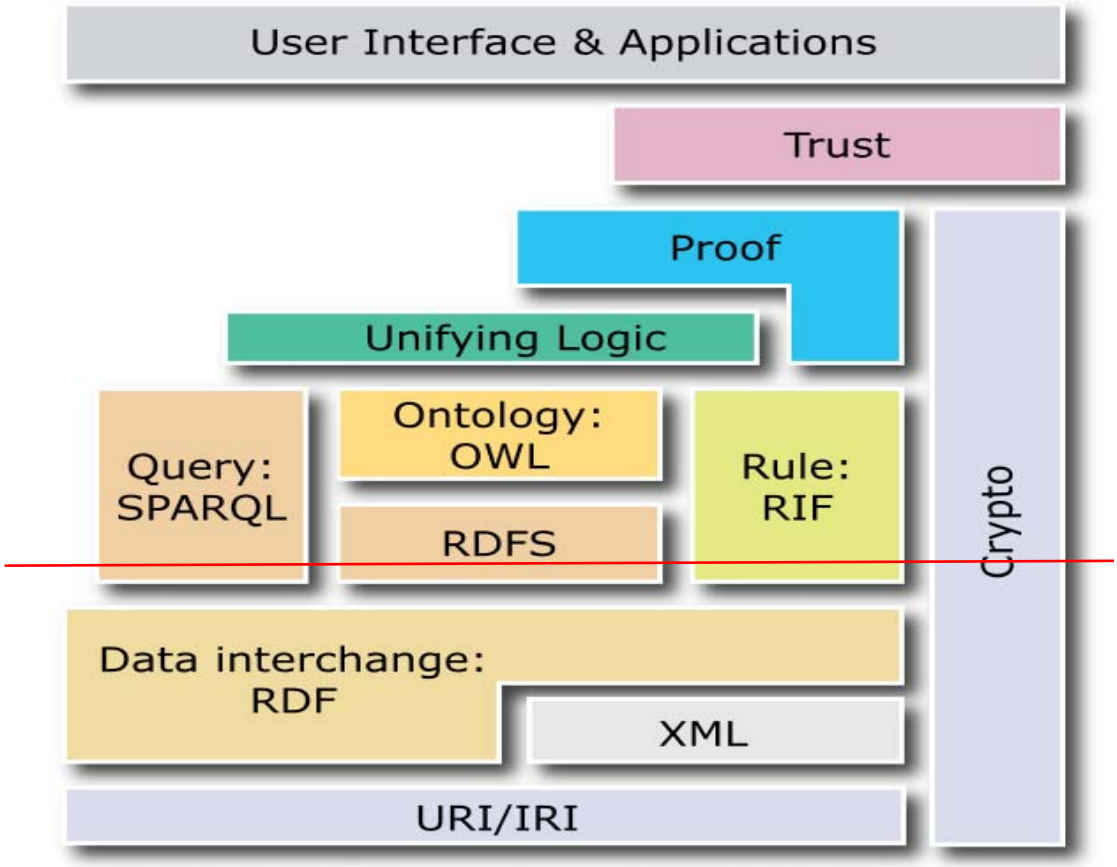

Abbildung 10: Semantisches Schichtenmodell [26]

Der Austausch von Informationen zwischen verschiedenen Anwendungen erfolgt über XML- bzw. RDF-basierte Sprachen. So können etwa Fakten, Aussagen, Anfragen mithilfe von RDF-Statements repräsentiert werden. Auf Basis eines gemeinsamen semantischen Modells kann eine übereinstimmende Interpretation der Bedeutung der entsprechenden Terme gewährleistet werden. RDF-S stellt bereits eine primitive Ontologiesprache dar, dennoch sind ausdrucksstärkere Sprachen, Darstellungsmöglichkeiten anwendungsspezifischen deklarativen Wissens durch Regelmengen und formale Logik und Mechanismen zur Validierung und Beweisführung erforderlich. Die Gewährleistung von Sicherheit und Qualität der verarbeiteten Informationen

erfordert eine weitere Ebene, die die Vertrauenswürdigkeit der Operationen sicherstellt.

Die Heterogenität der verteilten Informationsquellen im Web stellt die größte Herausforderung hinsichtlich deren Integration dar. Es müssen Konflikte auf syntaktischer, struktureller und semantischer Ebene aufgelöst werden. Der Integration auf semantischer Ebene kommt dabei eine besondere Bedeutung zu. Zum einen sind basierend auf den vorgestellten Standards und Datenbankanwendungen Integrationsansätze und -lösungen auf struktureller und syntaktischer Ebene bekannt und finden Verwendung, wobei aber keine inhaltliche Integration möglich ist. Zum anderen wird im Web vorrangig unstrukturierte Information vorgefunden, die einen semantischen Ansatz erforderlich macht. Dieser kann auf dem Zugang zur Semantik über die Struktur oder die natürlich-sprachliche Verarbeitung des Ursprungstextes basieren. Aufgrund der erwähnten Unstrukturiertheit vieler Web-Inhalte kommen vorrangig Methoden des Information Retrieval zum Einsatz, wobei komplexe linguistische Verfahren und Sammlungen von Wortschätzen sowie unterschiedlich komplexe semantische Modelle zur Darstellung von Kontext-Wissen und Konzeptionalisierungen Verwendung finden.

Je nach Grad der Ausdrucksstärke lassen sich verschiedene Ansätze unterscheiden, um Semantiken zu beschreiben und zu vergleichen. Sie können einfache Hierarchien von Ausdrücken oder komplexe Netzwerke sowie Hierarchien von Konzepten bis hin zu komplexen Konzeptbeschreibungen (Ontologien) umfassen. Die am häufigsten verwendeten Repräsentationen komplexer Konzepte sind framebasierte Darstellungen und Beschreibungslogiken. Vergleichsalgorithmen und logische Schließverfahren ermöglichen es, unterschiedliche Repräsentationen hinsichtlich ihres semantischen Inhalts zu vergleichen und ggf. zusammenzufassen.

Obwohl zu den verschiedenen Konflikttypen innerhalb der semantischen Integration Lösungsansätze erarbeitet und teils auch umgesetzt worden sind, kann man nicht davon sprechen, dass das Ziel eines semantischen Webs schon erreicht wäre. Einzelne isolierte Anwendungen vorrangig im universitären und wissenschaftlichen Umfeld demonstrieren zwar das Potential und verschiedene Werkzeuge und Quasi-Standards geben die Möglichkeit der Etablierung weiterer nützlicher Anwendungen. Dennoch sind trotz aller Fortschritte weiterhin sehr heterogene, spezialisierte und voneinander abgegrenzte Wissensbereiche einem semantischen Zugang geöffnet. Sowohl die Verwendung einheitlicher Ontologien als auch der Übergang von bestehenden darstellungsorientierten zu inhaltlich orientierten Informationsquellen, bedingt durch den Zugang zu deren Semantik, stellen mögliche Hindernisse dar.

Die Offenheit und Heterogenität des Webs, die die Notwendigkeit eines semantischen Ansatzes begründet haben, stellen somit auch die größte Herausforderung bei dessen Realisierung dar. So sind erste nennenswerte Erfolge am ehesten in Intranet-Umgebungen großer Organisationen, Universitäten oder Unternehmen zu erwarten [2].

Einige Kritiker bezweifeln überhaupt die Möglichkeit einer Etablierung eines semantischen Webs. Einerseits werden Zweifel hinsichtlich Annahmebereitschaft, Interessenlagen und Gewohnheiten der Nutzer [21, 22] und der Nutzbarkeit der derzeit im Web vorzufindenden Informationen [23], andererseits bezüglich des aktuellen Stands und Grenzen der Software-Entwicklung vorgebracht [24, 25].

Weitere Problemfelder öffnen sich durch Datenmissbrauch, Überwachung und Zensur. Die Notwendigkeit, in Ontologien Zusammenhänge der Welt zu kategorisieren, wirft die Frage auf, wer diese mit welchem ideologischen und kulturellen Hintergrund erzeugt und wie ein Konsens dazu gefunden werden kann.

3 Index

4 Literatur

[1] E. Rahm, G. Vossen (Hrsg.) Web & Datenbanken. dpunkt.verlag, 2003

[2] Antoniou, G. and Harmelen van, F.: *A Semantic Web Primer.* s.l. : The MIT Press, 2003.

[3] Hodgson, J.: *Do HTML Tags Flag Semantic Content.* In: IEEE INTERNET COMPUTING: 20-25, 01/02 2001.

[4] Stuckenschmidt, H. and Harmelen, van F.: *Information Sharing on the Semantic Web*. Springer-Verlag, Berlin Heidelberg New York, 2005.

[5] Bertino, E. and Ferrari, E.: *XML and Data Integration.* In: IEEE INTERNET COMPUTING: 75-76, 11/12 2001.

[6] Dörk: *Ontologiebasierte Informationsintegration*. Seminararbeit. Seminar: Texttechnologien und Semantic Web. Ernesto William De Luca .Otto-von-Guericke Universität Magdeburg, 2005.

[7] Boucelma, O., Castano, S., Goble, C., Josifovski, V., Lacroix, Z. and Ludächer, B.: *Report on the EDBT*. Panel on Scientific Data Integration. In: SIGMOD Record, 31(4): 107-112, December 2002.

[8] Doan, A. and Noy, A.: *Halevy Introduction to the Special Issue on Semantic Intgration*. In: SIGMOD Record, 33(4): 11-13, December 2004.

[9] Decker, S. Melnik, S. Harmelen, F. Van, Fensel, D., Klein, M., Broekstra, J., Erdmann, M., Horrocks, I.: *The Semantic Web: The Roles of XML and RDF*. In: IEEE INTERNET COMPUTING: 63-74, 09/10 2000.

[10] Panti, M., Spalazzi, L., Penzerini, L.: *Cooperation Strategies for Information Integration.* In: Cooperative Information Systems Proceedings of the 9th International Conference, CoopIS 2001, Trento, Italy, September 2001, Springer-Verlag, Berlin Heidelberg New York, 2001.

[11] Berlin, J. and Motro, A.: *Autoplex: Automated Discovery of Content for Virtual Databases.* In: Cooperative Information Systems Proceedings of the 9th International Conference, CoopIS 2001, Trento, Italy, September 2001, Springer-Verlag, Berlin Heidelberg New York, 2001.

[12] Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hübner, S.: *Ontology-Based Integration of Information- A Survey of Existing Approaches.* In: Ontology-based Information Integration

Tutorial at the 13th International Conference on Knowledge Engeneering and Knowledge Management (EKAW). Sigüenza, Spain, 01.-04.10.2002 Universität Bremen, FB Mathematik und Informatik, September 2002.

[13] Bergamaschi, S., Castano, S., Vincini, V.: *Semantic Integration of Semistructured and Structured Data Sources.* In SIGMOD Record, 28(1): 54-59, March 1999.

[14] Ouksel, A. M., Sheth, A.: *Semantic Interoperability in Global Information Systems A brief introduction to the research area and the special section*. In: SIGMOD Record, 28(1): 5-12, March 1999.

[15] U. Visser, U. and Stuckenschmidt, H.: *Interoperability in GIS-Enabling Technologies.* In: Ontology-based Information Integration Tutorial at the 13th International Conference on Knowledge Engeneering and Knowledge Management (EKAW). Sigüenza, Spain, 01.-04.10.2002 Universität Bremen, FB Mathematik und Informatik, September 2002.

[16] Stuckenschmidt, H. and I. Timm, I.: *Information Integration on the Semantik Web.* In: Ontology-based Information Integration Tutorial at the 13th International Conference on Knowledge Engeneering and Knowledge Management (EKAW). Sigüenza, Spain, 01.-04.10.2002 Universität Bremen, FB Mathematik und Informatik, September 2002.

[18] Patel, S. and Sheth, A.: *Semantic Information Requests Using Domain Modeling and Resource Characteristics.* In: Cooperative Information Systems Proceedings of the 9th International Conference, CoopIS 2001, Trento, Italy, September 2001, Springer-Verlag, Berlin Heidelberg New York, 2001.

[19] Tu, S. and Abdelguerfi, M.: *Web Services for Geographic Information Systems*. In: IEEE INTERNET COMPUTING: 12-14, 09/10 2006.

[20] G. Antoniou, G. and Boley, H. (eds*): Rules and RuleMarkup Languages for the Semantic Web.* Proceedings of the Third International Workshop RuleML 2004 Hiroshima, Japan, November 2004 Springer-Verlag, Berlin Heidelberg New York, 2004.

[21] Petrie, C.: *It's the Programming, Stupid*. In: IEEE INTERNET COMPUTING: 96-97, 05/06 2006.

[22] McCool, R.: *Rethinking the Semantic Web*, *Part II* In: IEEE INTERNET COMPUTING: 96-97, 01/02 2006.

[23] McCool, R.: *Rethinking the Semantic Web, Part I* In: IEEE INTERNET COMPUTING: 88-90, 11/12 2005.

[24] Petrie, C. and Bussler, C.: *Industrial Semantics and Magic.* In: IEEE INTERNET COMPUTING: 96-98, 07/08 2006.

[25] Hepp, M.: *Semantic Web and Semantic Web Services - Father and Son or Indivisible Twins?* In: IEEE INTERNET COMPUTING: 85-88, 03/04 2006.

[26] Berners-Lee, T., Hendler, J. and Lassila, O.:*The Semantic Web*. In: Scientific American Magazin: 35-44, May 2001.

[27] Sattler, K and Leymann, F.: *Information Integration & Semantic Web.* Datenbank-Spektrum: 5-6, 6/2003.