Präsentation

Thema **(Semi-)Automatisches Ontologie-Matching - *Teil 2***

Seminar Ontologiemanagement

Bearbeiter Thomas Efer

Betreuerin Sabine Maßmann

Abteilung am Datenbanken der Universität Leipzig
Institut für Informatik

Termin 13. Januar 2009

UNIVERSITÄT LEIPZIG
Abteilung Datenbanken
am Institut für Informatik

Was bisher geschah...

- Bedeutung von Ontologien
- Matching-Problem
- Matchansätze als „Werkzeugkiste“
- Wie geht es weiter?

UNIVERSITÄT LEIPZIG
Abteilung Datenbanken
am Institut für Informatik

Gliederung

Komposition

Problem: Einzelne Matchansätze haben jeweils ihre Stärken und Schwächen. Welche sind anzuwenden?

Idee: Verfolgung mehrerer Ansätze und sinnvolle Kombination der jeweiligen Ergebnisse

Komposition auf drei Ebenen denkbar:

- Built-In (statisch)
- Opportunistic (dynamisch)
- User-Driven (konfigurativ/interaktiv)

Komposition

Wie können die durch unterschiedliche Verfahren ermittelten Ähnlichkeitswerte fusioniert werden?

- $min(x_1, x_2, \dots)$
- $max(x_1, x_2, \dots)$
- $avg(x_1, x_2, \dots)$
- $\sqrt[k]{\frac{1}{n}\sum_{i=1}^{n}(x_i^k)}$
- $\dots$

Matrizenbasierte Optimierung

Wie kommt man von einer Ähnlichkeitsmatrix zu einem Alignment?

- Schwellwerte

Wie kommt man zu einem optimierten 1:1-Alignment?

- Betrachtung als Stable-Marriage-Problem
  („stabiles“ lokales Optimum, Beispiel folgt)
- Gewichtsmaximierung
  (globales Optimum, aufwändigere Berechnung)

UNIVERSITÄT LEIPZIG
Abteilung Datenbanken
am Institut für Informatik

Stable Marriage

S: ~~B~~, D, A, C
T: C, A, D, B
U: B, C, A, D
V: ~~D~~, A, C, B

A: T, S, V, U
B: V, U, S, T
C: S, V, U, T
D: T, S, V, U

- Sven macht Berta einen Heiratsantrag und sie stimmt unter Vorbehalt zu.
- Tom macht Carola einen Antrag und auch sie stimmt unter Vorbehalt zu.
- Ulf macht Berta einen Antrag. Sie stimmt unter Vorbehalt zu und lässt Sven sitzen. Armer Sven...
- Sven macht nun kurzerhand Diana einen Antrag. Sie stimmt unter Vorbehalt zu.
- Viktor macht Diana einen Antrag. Sie lacht ihn nur aus.
- Viktor lässt sich nicht entmutigen und macht nun Anna einen Antrag. Sie akzeptiert (unter Vorbehalt)

Graphenbasierte Strategien

Vorstellung am Beispiel des „Similarity Flooding“:

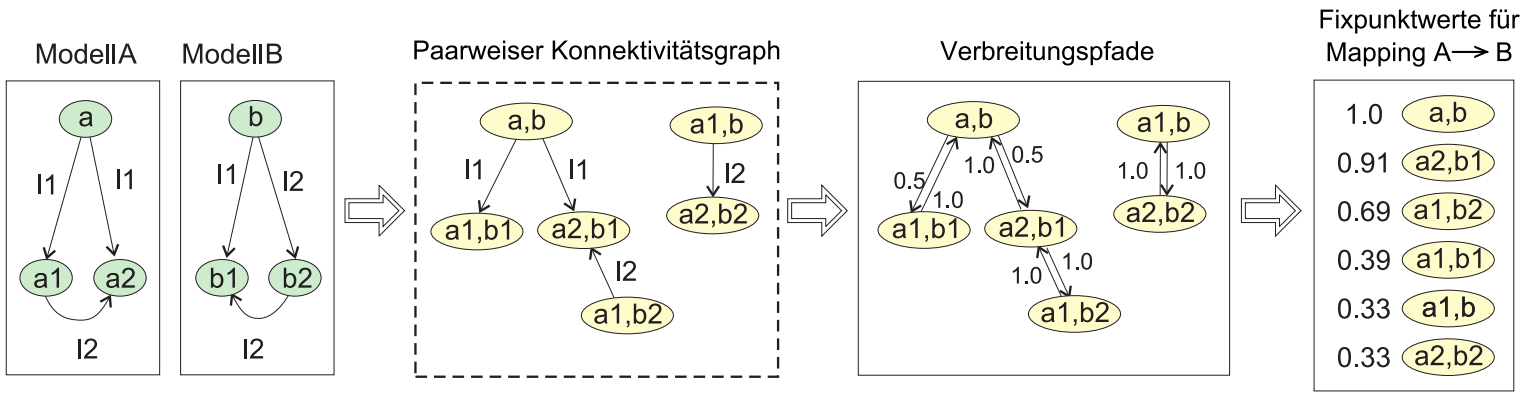


Abbildung: Similarity Flooding, Übersetzt aus [Melnik u. a., 2002]

UNIVERSITÄT LEIPZIG
Abteilung Datenbanken
am Institut für Informatik

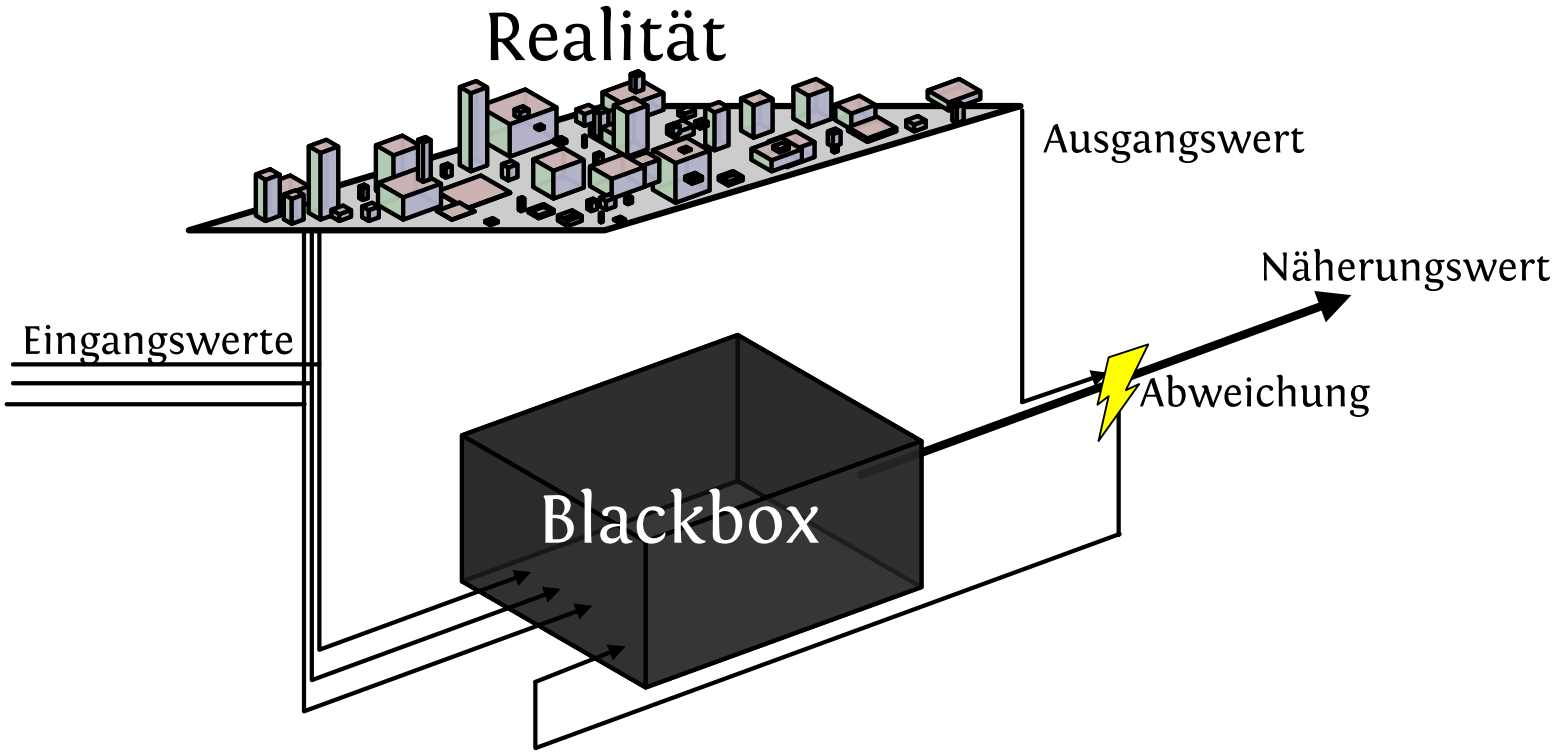
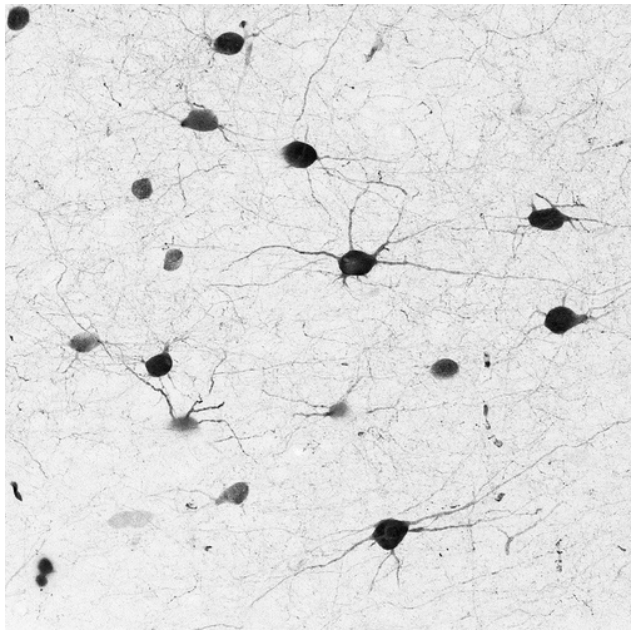
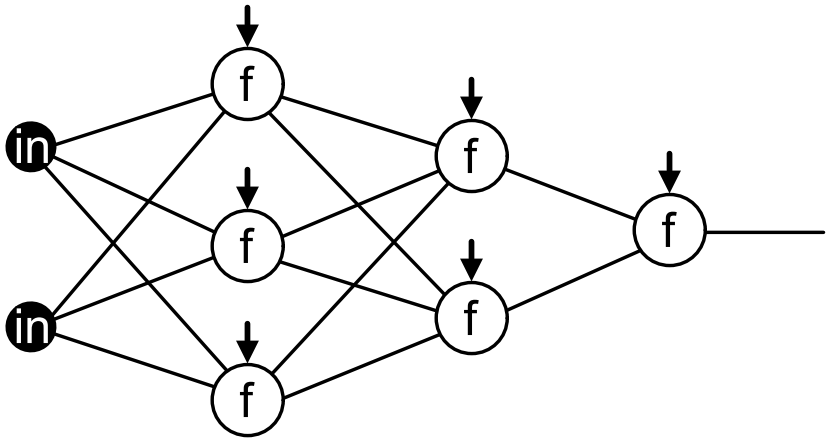


Abbildung: Grundlegender Ansatz „Trial and Error“

UNIVERSITÄT LEIPZIG
Abteilung Datenbanken
am Institut für Informatik

Künstliche Neuronale Netze





[http://people.brandeis.edu/~millerm/]

Bereits mehrere KNN-Ansätze veröffentlicht. [Huang u. a., 2007] [Mao u. a., 2008]

Genetische Algorithmen

Einführung verschiedener Analogien:

- DNA = Liste von $|O_1|$ Elementen, die jeweils $|O_2|$ Werte annehmen können = Correspondences
- Population = Liste von Individuen (mit eigener DNA)
- Paarung = Neues Individuum durch Crossover der DNAs (Kombination zuälliger DNA-Bereiche zweier Individuen)
- Mutation = zufällige Änderung von Teilen der DNA
- "Survival of the fittest" = Evaluierung des Alignments, das der DNA eines Individuums entspricht, Entfernen der ungeeignetsten

Die Evolution hilft, aus einer zufälligen Ursuppe hochentwickelte Individuen zu generieren. z.B. in [Wang u. a., 2006]

Überblick

Vielzahl von Matching-Werkzeugen vorhanden:

DELTA Hovy TranScm DIKE SKAT Artemis
H-Match Tess Anchor-Prompt OntoBuilder Cupid
COMA & COMA++ Similarity flooding XClust ToMAS
MapOnto OntoMerge CtxMatch S-Match HCONE
MoA ASCO BayesOWL OMEN DCM T-tree
CAIMAN FCA-merge LSD/GLUE/iMAP Automatch
SBI&NB KANG & Naughton Wang et al. sPI.Map
SEMINT Clio IF-ap NOM & QOM oMap Xu et al.
Wise-Integrator OLA Falcon-AO
Corpus-based-matching APFEL aflood aroma ASMOV
CIDER DSSim GeRoMe Lily MapPSO RiMOM
SAMBO SPIDER TaxoMap ...

Überblick

Unterschiedlicher Fokus der einzelnen Tools, beispielsweise:

- Plugins für Ontologieeditoren
- Tools aus der Schemaverwaltung
- Standalone Prototypen für Demonstration einzelner Algorithmen

Nachfolgend wird ein Querschnitt über häufige Features aller Produkte vorgestellt.

Eingabeformate

Quelle für Schema- und Instanzdaten:

- Relationale Datenbanken und deren Schema; Tripeldatenbanken
- Generische XML-Formate und deren XSD beziehungsweise DTD
- Ontologie-Formate wie OWL (in Form von RDF oder XML)

UNIVERSITÄT LEIPZIG
Abteilung Datenbanken
am Institut für Informatik

Interaktion

- Interaktion über GUI, eventuell mit graphischer Repräsentation der Ontologien oder des aktuellen Alignments
- Programmatischer Zugriff auf die Funktionen über Schnittstellen (APIs)
- Verschiedene Formen der Nutzerinteraktion:
  - Auswahl von Matchansätzen
  - Auswahl von Untermengen der Ontologien für Teilmatching
  - Refinement von Alignments

Funktionsprinzipien

Komposition: Vollständige Analyse mit mehreren Ansätzen oder hierarchischer Ansatz zur Verfeinerung

Struktur: Pfadverfolgung weniger wichtig als in Schemamatching, zum Teil werden fragmentbasierte Ansätze verwendet

Instanzen: Meist zunächst Matching der Konzepte ohne Instanzen, danach Verifikation über Instanzdaten

Feature-Querschnitt

externe Quellen

- Thesauri
- Domänenspezifische Vokabulare
- WordNet (`http://wordnet.princeton.edu/`)

Ausgabeformen

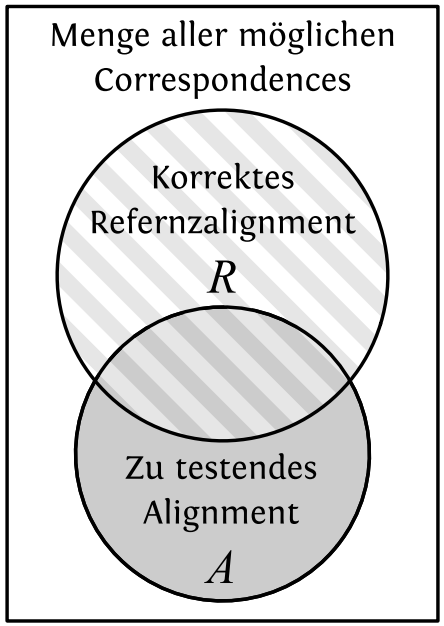
- Alignments zum Beispiel in Tripel-Darstellung
- Views als Verknüpfung relationaler Daten
- Regel-Sätze

UNIVERSITÄT LEIPZIG
Abteilung Datenbanken
am Institut für Informatik

mit Referenz

Übliche Vorgehensweise:

- Manuelle Ermittlung eines Referenzalignments R
- Vergleichen des zu untersuchenden Alignments A mit R



UNIVERSITÄT LEIPZIG
Abteilung Datenbanken
am Institut für Informatik

Dieser Vergleich kann unter verschiedenen Gesichtspunkten ablaufen...

Ausgangsbasis: Numerische Betrachtung der Alignments als Menge von Correspondences.

Verschiedene Maße eingeführt:

- Der *Hammingabstand* als „Unähnlichkeitsmaß“ zwischen Refernzalignment R und dem zu testenden Alignment A

$$H(A, R) = 1 - \frac{|A \cap R|}{|A \cup R|}$$

mit Referenz

- Die *Precision* ist das Verhältnis der korrekt gefundenen Correspondences zu allen Ermittelten (inklusive „falscher Positiver“).

$$Precision(A, R) = \frac{|A \cap R|}{|A|}$$

- Der *Recall* setzt die korrekt gefunden Correspondences ins Verhältnis zu allen korrekten Referenzalignments setzt.

$$Recall(A, R) = \frac{|A \cap R|}{|R|}$$

UNIVERSITÄT LEIPZIG
Abteilung Datenbanken
am Institut für Informatik

- Die *F-Measures* erlauben eine Einbeziehung von Precision und Recall in einen einzigen Wert, wobei zwischen beiden gewichtet harmonisch gemittelt wird. ($\alpha \in [0; 1]$)

$$F_\alpha = \frac{Precision \cdot Recall}{(1-\alpha) \cdot Precision + \alpha \cdot Recall}$$

- Der *Overall* beziehungsweise die *Accuracy* kombiniert beide Werte auf eine andere Art und Weise:

$$Overall = Recall \cdot (2 - Precision^{-1})$$

ohne Referenz

In Real-Life-Situationen:

- Bildung eines kompletten Referenzalignments ist nicht sinnvoll
- Extrapolation nach Bildung nur einer kleinen Teilreferenz ist ungenau und dennoch mit Aufwand verbunden
- Maß ohne R wird benötigt, wobei bestimmte Annahmen über ein gutes Alignment getroffen werden müssen, z.B. das Vorhandensein vieler 1:1-Beziehungen

ohne Referenz

Ansatz nach [Kirsten u. a., 2007]:

- Die *MatchCoverage* einer der zu matchenden Ontologien als Quotient aus gematchten Konzepten zu allen Konzepten:

$$MC_{O_1} = \frac{|O_{1Matched}|}{|O_1|} \qquad (O_2 : analog)$$

- Bei der *InstanceMatchCoverage* ist der Divisor die Anzahl der Konzepte mit wenigstens einer zugeordneten Instanz.

$$IMC_{O_1} = \frac{|O_{1Matched}|}{|O_{1Inst}|} \qquad (O_2 : analog)$$

- Die *MatchRatio* setzt die gefundenen Correspondenzes in Relation zur Anzahl gematchter Konzepte.

$$MR_{A,O_1} = \frac{|A|}{|O_1|} \qquad (O_2 : analog)$$

Status

- Matchingqualität und Performance aktueller Systeme sind bereits ziemlich weit fortgeschritten und erleichtern Alignmentaufgaben kleinen und mittleren Umfangs.
- Geeignete Evaluierungsmethoden sind bekannt. Auch ein internationaler Wettbewerb zur Evaluierung neuer Methoden existiert (OAEI: `http://oaei.ontologymatching.org`)
- Dennoch gibt es einen viel größeren Bedarf an frei zugänglichen Benchmarks.

UNIVERSITÄT LEIPZIG
Abteilung Datenbanken
am Institut für Informatik

Weiterentwicklung

- Schnellere Algorithmen mit „besseren“ Alignments
- Bessere Skalierbarkeit für das Matching sehr großer Ontologien
- Aufbereitung weiterer Quellen für benötigtes Domänenwissen
- Geeignete Visualisierungstechniken für bessere Interaktivität [Kotis u. Lanzenberger, 2008]

Anschließende Arbeitsschritte

Für Merging ist es sinnvoll, zunächst mehrere Alignments zu erstellen und gemeinsam zu verwalten (Mapping-Management).

Alternativ müssen für Datentransformation oder -integration, sogenannte „Mapping Expressions“ für die Transformation der Instanzdaten gefunden werden. (Typkonvertierungen, Umrechnungsoperationen, Zusammenfügen und Splitten von Zeichenketten, etc.)

Zusätzlich sind vorberechnete Alignments auch als Basis für eine effiziente Abarbeitung von Queries an heterogen zusammengesetzte Wissensbasen geeignet.

[Huang u. a. 2007] Huang, Jingshan ; Dang, Jiangbo ; Vidal, José M. ; Huhns, Michael N.: Ontology Matching Using an Artificial Neural Network to Learn Weights. In: *IJCAI Workshop on Semantic Web for Collaborative Knowledge Acquisition*, 2007

[Kirsten u. a. 2007] Kirsten, Toralf ; Thor, Andreas ; Rahm, Erhard: Instance-based matching of large life science ontologies. In: *Data Integration in the Life Sciences*, 2007

[Kotis u. Lanzenberger 2008] Kotis, Konstantinos ; Lanzenberger, Monika: Ontology Matching: current status, dilemmas and future challenges. In: *International Conference on Complex, Intelligent and Software Intensive Systems*, 2008

[Mao u. a. 2008] Mao, Ming ; Peng, Yefei ; Spring, Michael: Integrating the IAC Neural Network in Ontology Mapping. In: *17th International World Wide Web Conference*, 2008

[Melnik u. a. 2002] Melnik, Sergey ; Garcia-Molina, Hector ; Rahm, Erhard: Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching. In: *18th International Conference on Data Engineering*, 2002

[Wang u. a. 2006] Wang, Junli ; Ding, Zhijun ; Jiang, Changjun: GAOM: Genetic Algorithm based Ontology Matching. In: *IEEE Asia-Pacific Conference on Services Computing*, 2006