Learning Expressive Linkage Rules using Genetic Programming

R. Isele and C. Bizer
Seminarvortrag von Paul Dubs

TECHNISCHE UNIVERSITÄT DARMSTADT

Einleitung

TECHNISCHE UNIVERSITÄT DARMSTADT

- Problem Definition
- Kurzer Überblick über Genetic Programming
- GenLink
  - Darstellung von Linkage Rule als Baum
  - Vorverarbeitung der Daten
  - Fitness Funktion
  - Crossover Operationen
- Alternative Ansätze
- Experimentelle Evaluierung
  - Ergebnisse auf Unterschiedlichen Datensätzen
  - Einfluss von einzelnen Features
- Kritik & Ausblick

Linkage Rules

TECHNISCHE UNIVERSITÄT DARMSTADT

- Regeln zum Feststellen ob zwei Objekte in zwei unterschiedlichen Datenbanken das gleiche reelle Objekt darstellen
- Nützlich für...
  - … Verbesserung der Datenqualität in einer Datenbank
  - … Erkennen der gleichen reellen Objekte in unterschiedlichen Datenbanken mit unterschiedlichem Schema
    - z.B. beim Zusammenführen von Datenbanken bei Firmenübernahme oder Zusammenführen von Informationen aus dem Semantic Web
- Manuelles erstellen von Linkage Rules braucht Expertenwissen da zusammenhänge häufig auch nicht Trivial sein können

Genetic Programming

TECHNISCHE UNIVERSITÄT DARMSTADT

- Genetic Programming ist eine Spezialisierung von Genetic Algorithms
- Jedes Individuum in der Population ist dabei ein Programm
- Programme werden als Baum dargestellt
- Gene sind Knoten und Blätter des Baumes
- Genaue Bedeutung der Gene ist abhängig von Problemstellung
- Fitnessfunktion weißt jedem Individuum einen Wert zu der Aussagt wie gut dieses das Problem löst
- Durch wiederholtes kreuzen und mutieren von den fittesten Individuen soll eine möglichst gute Lösung gefunden werden
- Nicht zwangsläufig deterministisch

GenLink

TECHNISCHE UNIVERSITÄT DARMSTADT

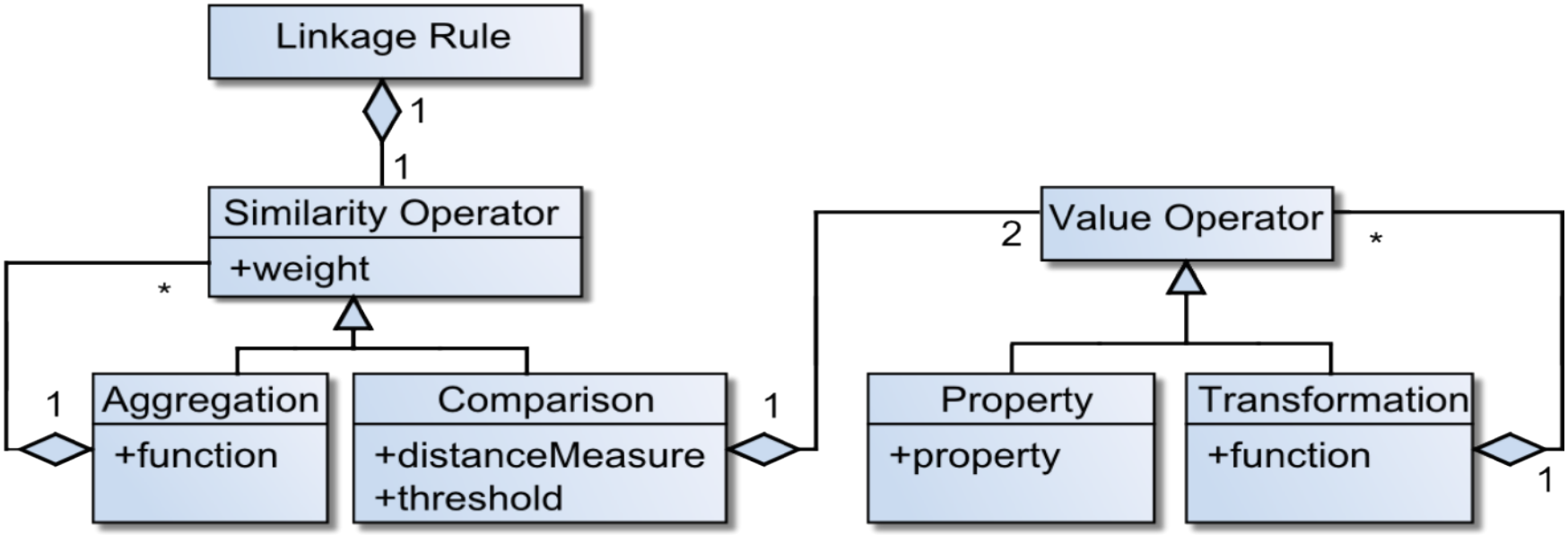
- Unterstützt Transformationen
- Präferiert kurze Linkage Rules
- Benötigt Referenz Links (positive und negative)
- Vorverarbeitung von Referenz Links für bessere Ausgangspopulation
- Feste Regeln für Operator-Komposition
- Intelligente Crossover Funktion anstatt Subtree Crossover
- Mutation durch Headless Chicken Crossover
- Matthews Correlation Coefficient als primäre Fitnessfunktion
- Tournament Selection

GenLink Operatoren

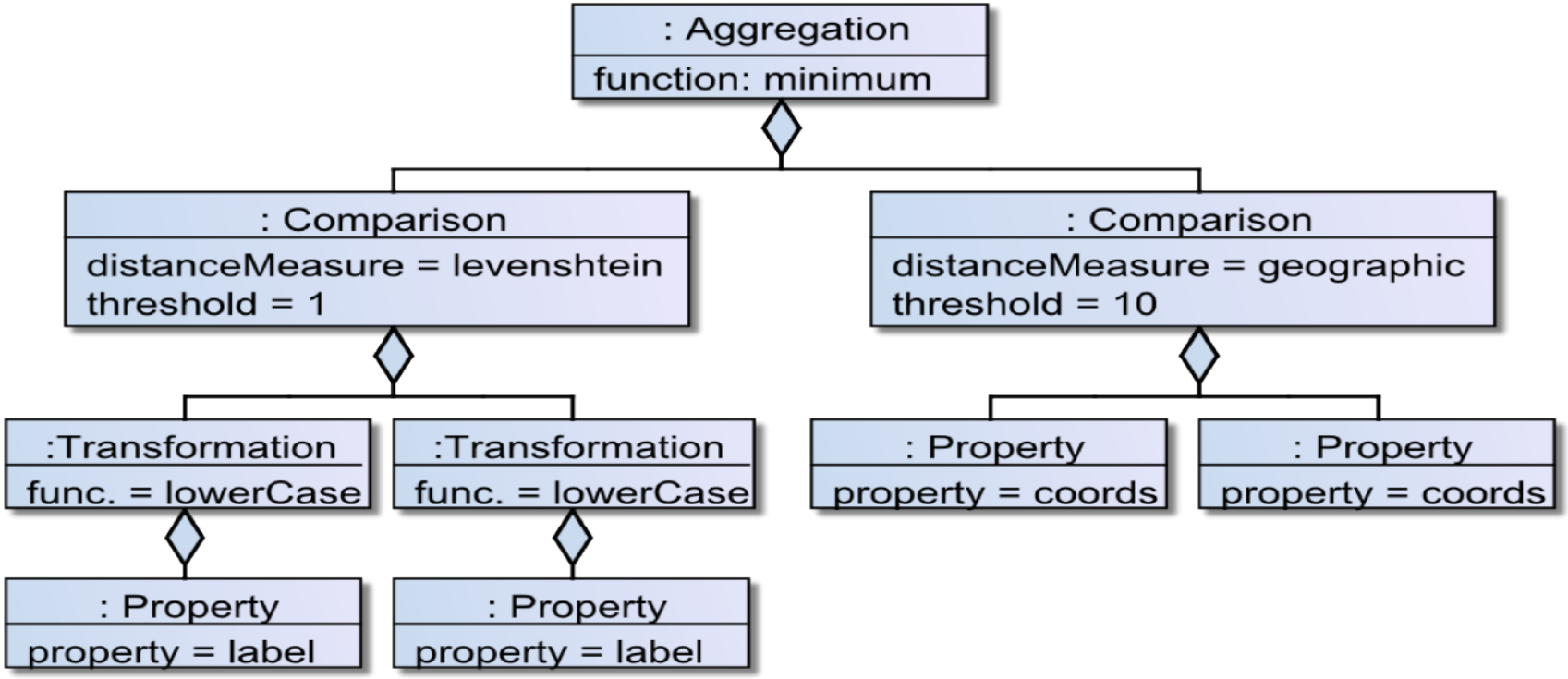
TECHNISCHE UNIVERSITÄT DARMSTADT

- **Property:**
  - Gibt einen Wert des Objektes zurück
- **Transformation:**
  - Transformiert einen Wert anhand einer Transformierungsfunktion
- **Comparison:**
  - Vergleicht Werte anhand eines Ähnlichkeitsmaßes und eines Grenzwertes
- **Aggregation:**
  - Aggregiert die Werte von mehreren Comparison und Aggregation Operatoren anhand einer Aggregierungsfunktion
- **Similarity:**
  - Gruppierung von Comparison und Aggregation Operatoren
  - Gibt eine Gewichtung vor

GenLink Operator Struktur



GenLink Linkage Rule Beispiel



Vorverarbeitung für bessere initial Population

TECHNISCHE UNIVERSITÄT DARMSTADT

- Suchraum kann sehr Groß werden
  - Wenn Objekte sehr viele Eigenschaften haben
  - Objekte aus Datensätzen stammen welche unterschiedliche Schemas verwenden
- Analyse von positiven Referenz Links zur Erstellung einer Liste von Eigenschaftspaaren
- Generierung von Zufallsregeln greift auf diese Liste zurück um bessere Ausgangspunkte zu finden

Erstellen der Eigenschaftspaar-Liste

TECHNISCHE UNIVERSITÄT DARMSTADT

- Gehe alle Eigenschaftspaare von positiven Referenz Links durch
- Wandele Eigenschaftswert jeweils in Kleinschreibung um
- Tokenifizere den kleingeschriebenen Wert
- Falls die so erlangten Werte anhand einer Ähnlichkeitsmetrik innerhalb eines Grenzwertes liegen, füge das Eigenschaftspaar zur Liste hinzu

Generieren einer zufälligen Linkage Rule

TECHNISCHE UNIVERSITÄT DARMSTADT

- Zufällig generierte Linkage Rules beinhalten einen zufällig gewählten Aggregation Operator und bis zu maximal zwei Comparison Operatoren
- Jeder Comparison Operator wählt für seine Eingabewerte ein Eigenschaftspaar aus der vor generierten Liste
- Mit einer 50% Wahrscheinlichkeit wird dabei auch eine Transformation auf dieser Eigenschaft durchgeführt

Auswahl der besten Regel

TECHNISCHE UNIVERSITÄT DARMSTADT

- Matthews Correlation Coefficient ist größter Teil der Fitnessfunktion

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)\cdot(TP+FN)\cdot(TN+FP)\cdot(TN+FN)}}$$

- Kurze Regeln sollen bevorzugt werden deswegen spielt Operator Anzahl negativ mit hinein

$$fitness = MCC - 0.05 \cdot operatorCount$$

Intelligentes Crossover

TECHNISCHE UNIVERSITÄT DARMSTADT

- Verwendet Menge von Crossover Funktionen die an Linkage Rule Struktur angepasst sind
- Jedes Crossover wählt zufällig eine dieser Funktionen aus
- Mutation durch Crossover mit zufällig generierter Linkage Rule
- Crossover arbeitet auf jeweils 2 Linkage Rules
  - In den meisten Fällen gibt es einen Dominanten und einen Rezessiven Part
  - Dominanter Part übergibt das Meiste seiner Regeln
  - Rezessiver Part trägt nur einen kleinen Teil bei

Crossover Funktionen

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Function Crossover:**
  - Findet die beste Distanz, Transformation oder Aggregationsfunktion
  - Wählt jeweils zufällige Comparison, Transformation oder Aggregation Operatoren und tauscht diese
- **Operators Crossover:**
  - Kombiniert zwei Aggregation Operatoren
  - Modifiziert eine Aggregation des Dominanten Parts
  - Wählt zufällig jeweils eine Aggregation, fügt im Dominanten Part alle Comparison Operatoren des Rezessiven Parts zur Aggregation hinzu, geht dann alle Comparison Operatoren dieser Aggregation durch und löscht mit einer Wahrscheinlichkeit von 50% diesen Operator

Crossover Funktionen

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Aggregation Crossover:**
  - Baut Hierarchien auf
  - Wählt jeweils zufällig Aggregation Operator oder Comparison Operator und ersetzt den gewählten Part im Dominanten durch den Rezessiven
  - Ähnlich zu Subtree Crossover, arbeitet aber nur auf Aggregation und Comparison Operatoren
- **Transformation Crossover:**
  - Kombiniert Transformationen, kann Transformationsketten erstellen
  - Wählt jeweils zufällig zwei Transformation Operatoren, sodass einer im Baum über dem anderen liegt und vertauscht den dazwischen liegenden Pfad durch Two Point Crossover

Crossover Funktionen

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Threshold Crossover:**
  - Kombiniert Differenz Grenzwerte
  - Wählt jeweils einen Comparison Operator und setzt den Grenzwert im Dominanten Part auf den Durchschnitt der beiden gewählten
- **Weight Crossover:**
  - Kombiniert Gewichtungen
  - Wählt jeweils einen Comparison Operator oder Aggregation Operator und setzt das Gewicht im Dominanten Part auf den Durchschnitt der beiden gewählten

Alternative Ansätze

TECHNISCHE UNIVERSITÄT DARMSTADT

- **A Genetic Programming Approach to Record Deduplication** von Carvalho et al. (2011)
  - Genbaum beschreibt lineare Kombinationen von Ähnlichkeitsergebnissen
  - F1-Measure als Fitnessfunktion
- **Naive Bayes**
  - Schlechtere Performance als Lineare Klassifizierer und Entscheidungsbäume
- **Lineare Klassifizierer**
  - z.B. MARLIN
  - Verwenden häufig SVMs
- **Grenzwert-basierte Bool'sche Klassifizerer**
  - Verwenden typischer weise Entscheidungsbäume
- Alle genannten unterstützen in der Regel keine Transformationen

Experimentelle Evaluierung

TECHNISCHE UNIVERSITÄT DARMSTADT

- Implementiert im Silk Link Discovery Framework
- Alle Experimente mit GenLink jeweils 10 mal ausgeführt
  - Datensatz jeweils zufällig in Trainingset und Testset aufgeteilt
  - Ergebnisse wurden gemittelt und Standardabweichung wurde berechnet
- Experimente liefen auf 3GHz Intel Core i7 (4 Cores)
- Java Heap Space limitiert auf 1GB
- Kein Dataset liefert Negative Referenz Links mit aus, daher wurden diese generiert

GenLink Parameter

TECHNISCHE UNIVERSITÄT DARMSTADT

Parameter	Value
Population size	500
Maximum iterations	50
Selection method	Tournament selection
Tournament size	5
Probability of Crossover	75%
Probability of Mutation	25%
Stop Condition	F-measure = 1.0

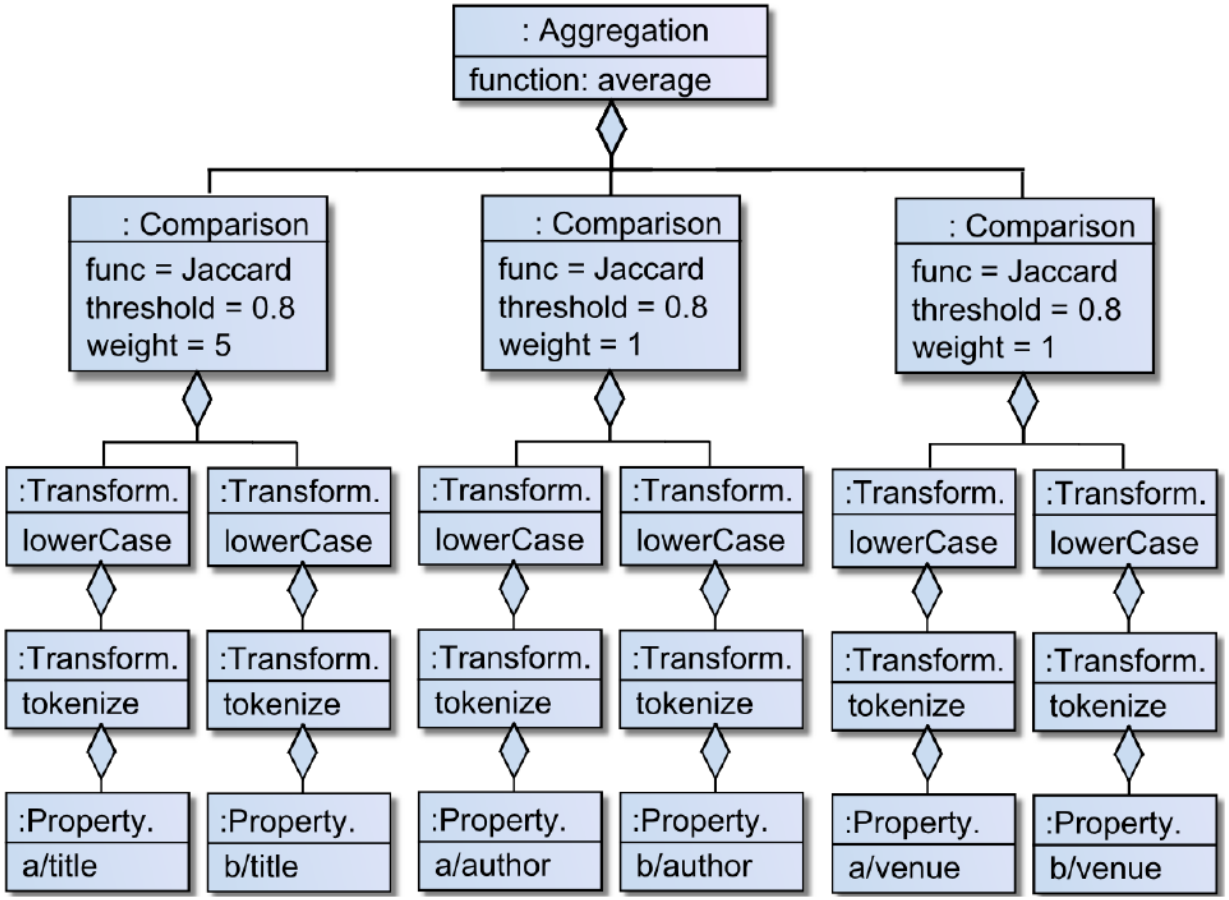
Cora

TECHNISCHE UNIVERSITÄT DARMSTADT

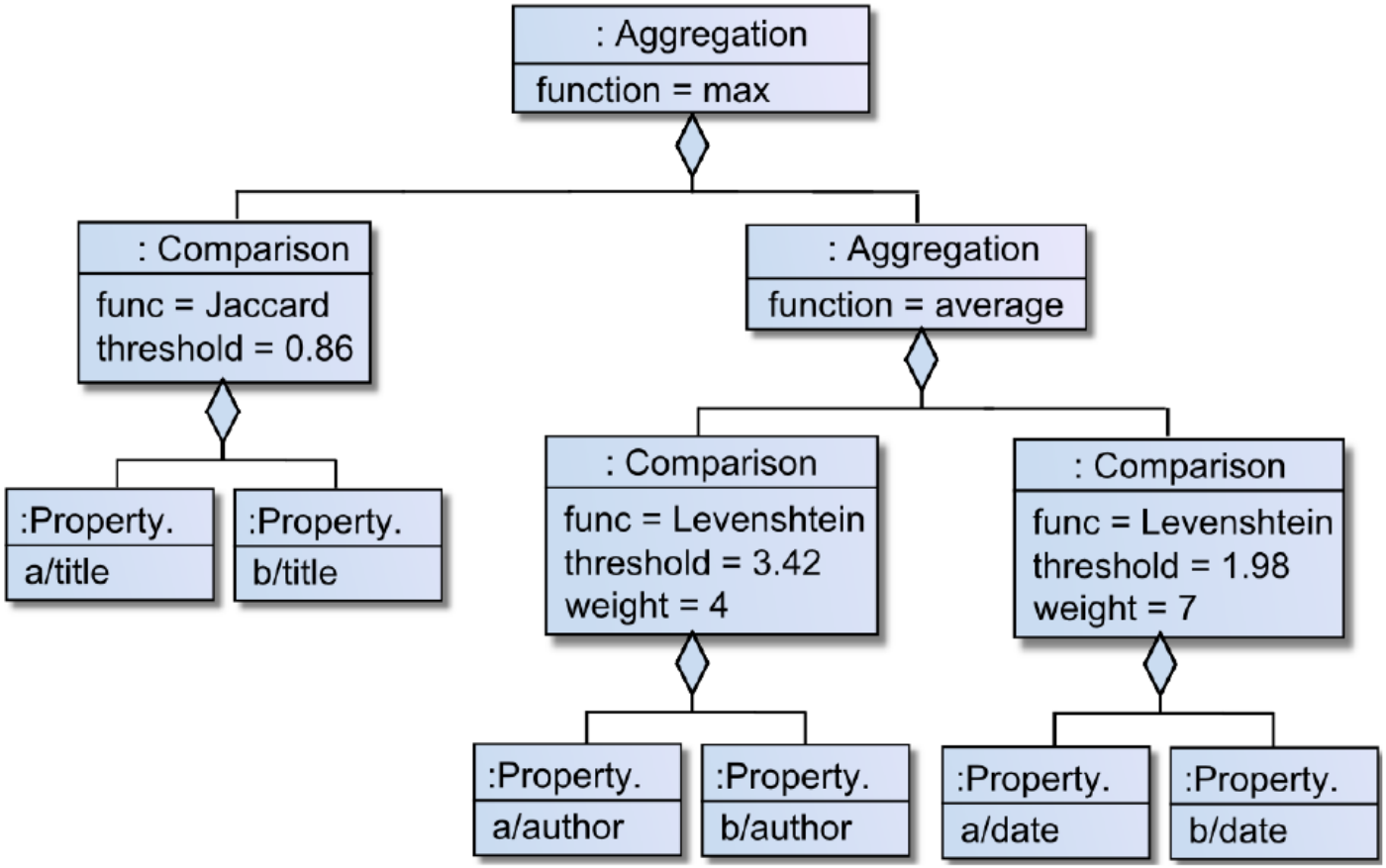
- Beinhaltet Zitat Quellen aus der Cora Compter Science Research Paper Search Engine
- Für jeden Datensatz gibt es den Titel, Author, Veröffentlichungsort und Datum
- Referenz Wert von Carvalho et al.

Iter.	Time in s (σ)	Train. F1 (σ)	Val. F1 (σ)
0	5.5 (0.7)	0.880 (0.030)	0.877 (0.031)
10	28.6 (2.7)	0.949 (0.018)	0.945 (0.021)
20	60.1 (4.1)	0.965 (0.005)	0.962 (0.005)
30	93.6 (6.1)	0.968 (0.003)	0.965 (0.004)
40	129.4 (9.7)	0.968 (0.002)	0.965 (0.004)
50	185.8 (26.7)	0.969 (0.003)	0.966 (0.004)
Ref.	-	0.900 (0.010)	0.910 (0.010)

Cora Linkage Rule



Cora Linkage Rule ohne Transformationen



Restaurant

TECHNISCHE UNIVERSITÄT DARMSTADT

- Restaurant Datensätze aus Fodor's und Zagat's Restaurant Führern
- Jeder Eintrag beinhaltet Name, Adresse, Telefonnummer und Art des Restaurants
- Referenz Wert von Carvalho et al.

Iter.	Time in s (σ)	Train. F1 (σ)	Val. F1 (σ)
0	0.4 (0.1)	0.953 (0.038)	0.951 (0.039)
10	2.0 (0.9)	0.996 (0.004)	0.992 (0.006)
20	3.1 (1.9)	0.996 (0.004)	0.993 (0.006)
30	4.1 (3.0)	0.996 (0.004)	0.993 (0.006)
40	5.2 (4.0)	0.996 (0.004)	0.993 (0.006)
50	6.3 (5.3)	0.996 (0.004)	0.993 (0.006)
Ref.	-	1.000 (0.000)	0.980 (0.010)

Sider – DrugBank

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ontology Alignment Evaluation Initiative (OAEI) Datensatz von 2010
- Konkurrenz ist verwendet keine Referenz Links (ist unsupervised), daher als Baseline anzusehen
- Datensatz über Medikamente und ihre Nebenwirkungen, beinhaltet Daten von Sider und DrugBank

Iter.	Time in s (σ)	Train. F1 (σ)	Val. F1 (σ)
0	21.7 (0.3)	0.840 (0.018)	0.837 (0.018)
10	38.8 (2.4)	0.943 (0.025)	0.939 (0.030)
20	83.1 (11.1)	0.970 (0.007)	0.969 (0.008)
30	147.2 (20.9)	0.972 (0.006)	0.970 (0.007)
40	215.6 (28.0)	0.972 (0.006)	0.970 (0.007)
50	301.5 (39.0)	0.972 (0.006)	0.970 (0.007)

Reference System	F1
ObjectCoref	0.464 [18]
RiMOM	0.504 [30]

NYT Datensatz

TECHNISCHE UNIVERSITÄT DARMSTADT

- Orte in New York Times und DBPedia
- Aus OAEI 2011

Iter.	Time in s (σ)	Train. F1 (σ)	Val. F1 (σ)
0	85.2 (1.7)	0.703 (0.064)	0.709 (0.048)
1	107.7 (11.0)	0.803 (0.037)	0.803 (0.036)
5	260.7 (76.8)	0.844 (0.048)	0.846 (0.048)
10	344.5 (86.2)	0.854 (0.052)	0.854 (0.053)
20	496.7 (95.0)	0.907 (0.074)	0.906 (0.074)
30	652.8 (108.1)	0.927 (0.069)	0.928 (0.067)
40	804.8 (132.4)	0.965 (0.039)	0.963 (0.041)
50	975.4 (141.1)	0.977 (0.024)	0.974 (0.026)

Reference System	F1 [13]
AgreementMaker	0.69
SEREMI	0.68
Zhishi.links	0.92

Linked MDB

TECHNISCHE UNIVERSITÄT DARMSTADT

- Film Datensatz der nicht-triviale Linkage Rules braucht
- Kein Vergleichwert
- Produzierte Linkage Rule ist die gleiche die ein Mensch auch anlegen würde

Iter.	Time in s (σ)	Train. F1 (σ)	Val. F1 (σ)
1	4.2 (1.2)	0.981 (0.011)	0.959 (0.023)
10	19.6 (2.1)	0.998 (0.001)	0.921 (0.007)
20	40.3 (3.3)	1.000 (0.000)	0.974 (0.004)
30	59.2 (4.2)	1.000 (0.000)	0.999 (0.002)
40	74.2 (7.2)	1.000 (0.000)	0.999 (0.002)
50	99.7 (12.8)	1.000 (0.000)	0.999 (0.002)

DBPedia - DrugBank

TECHNISCHE UNIVERSITÄT DARMSTADT

- Benötigt eine Komplexe Linkage Rule weil häufig nicht alle Daten für jedes Objekt verfügbar sind
- Manuell erstellte Linkage Rule braucht 13 Vergleiche und 33 Transformationen
- Durch GenLink erstellte Linkage Rule braucht im Durchschnitt 5.6 Vergleiche und 3.2 Transformationen

Iter.	Time in s (σ)	Train. F1 (σ)	Val. F1 (σ)
1	67.5 (2.2)	0.929 (0.026)	0.928 (0.029)
10	334.1 (157.4)	0.994 (0.002)	0.991 (0.003)
20	1014.1 (496.8)	0.996 (0.001)	0.988 (0.010)
30	1829.7 (919.3)	0.997 (0.001)	0.985 (0.016)
40	2685.4 (1318.9)	0.998 (0.001)	0.994 (0.002)
50	3222.2 (1577.7)	0.998 (0.001)	0.994 (0.002)

Einfluss von Teilfeatures

TECHNISCHE UNIVERSITÄT DARMSTADT

	Boolean	Linear	Nonlin.	Full
Cora	0.900	0.896	0.898	0.965
Restaurant	0.954	0.959	0.951	0.992
SiderDrugBank	0.931	0.956	0.966	0.970
NYT	0.714	0.716	0.724	0.916
LinkedMDB	0.973	0.986	0.987	0.997
DBpediaDrugBank	0.990	0.981	0.991	0.993

	Random	Seeded
Cora	0.849 (0.045)	0.865 (0.018)
Restaurant	0.963 (0.010)	0.985 (0.012)
SiderDrugBank	0.624 (0.181)	0.848 (0.013)
NYT	0.178 (0.164)	0.701 (0.072)
LinkedMDB	0.719 (0.175)	0.975 (0.008)
DBpediaDrugBank	0.702 (0.217)	0.957 (0.013)

25 Iterations	Subtree C.	Our Approach
Cora	0.959 (0.007)	0.967 (0.003)
Restaurant	0.997 (0.004)	0.997 (0.004)
SiderDrugBank	0.974 (0.004)	0.987 (0.003)
NYT	0.814 (0.005)	0.916 (0.006)
LinkedMDB	0.996 (0.007)	0.998 (0.003)
DBpediaDrugBank	0.994 (0.001)	0.997 (0.002)

- Vergleich von unterschiedlichen Linkage Rule Representierungen
- Auswirkungen der Vorverarbeitung
- Intelligente Crossover Funktion

Kritik & Ausblick

TECHNISCHE UNIVERSITÄT DARMSTADT

- GenLink ist angeblich im Silk Link Discovery Framework implementiert, leider konnte ich die Implementierung darin nicht finden
- Das Paper ist eine etwas formalisierte Version eines Papers das die Autoren auf einem Workshop vorgestellt haben
- Die gleiche Idee wurde im Zusammenhang mit der Minimierung von den Notwendigen Referenz Links von den Autoren schon mal vorgestellt
- Vorgehensweise ist gut beschrieben
- Das Workshop Paper wurde von anderen Autoren auch aufgegriffen

Bibliography

- R. Isele and C. Bizer. Learning expressive linkage rules using genetic programming. Proceedings of the VLDB Endowment, 5(11), pages 1638-1649, 2012.
- R. Isele and C. Bizer. Learning linkage rules using genetic programming. In 6th International Workshop on Ontology Matching, pages 13–24,2011.
- M. Carvalho, A. Laender, M. Gon¸alves, c and A. da Silva. Replica identification using genetic programming. In Proceedings of the 23rd Annual ACM Symposium on Applied Computing, pages 1801–1806,2008.

Fragen?

TECHNISCHE UNIVERSITÄT DARMSTADT