OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

INF FAKULTÄT FÜR INFORMATIK

DBSE

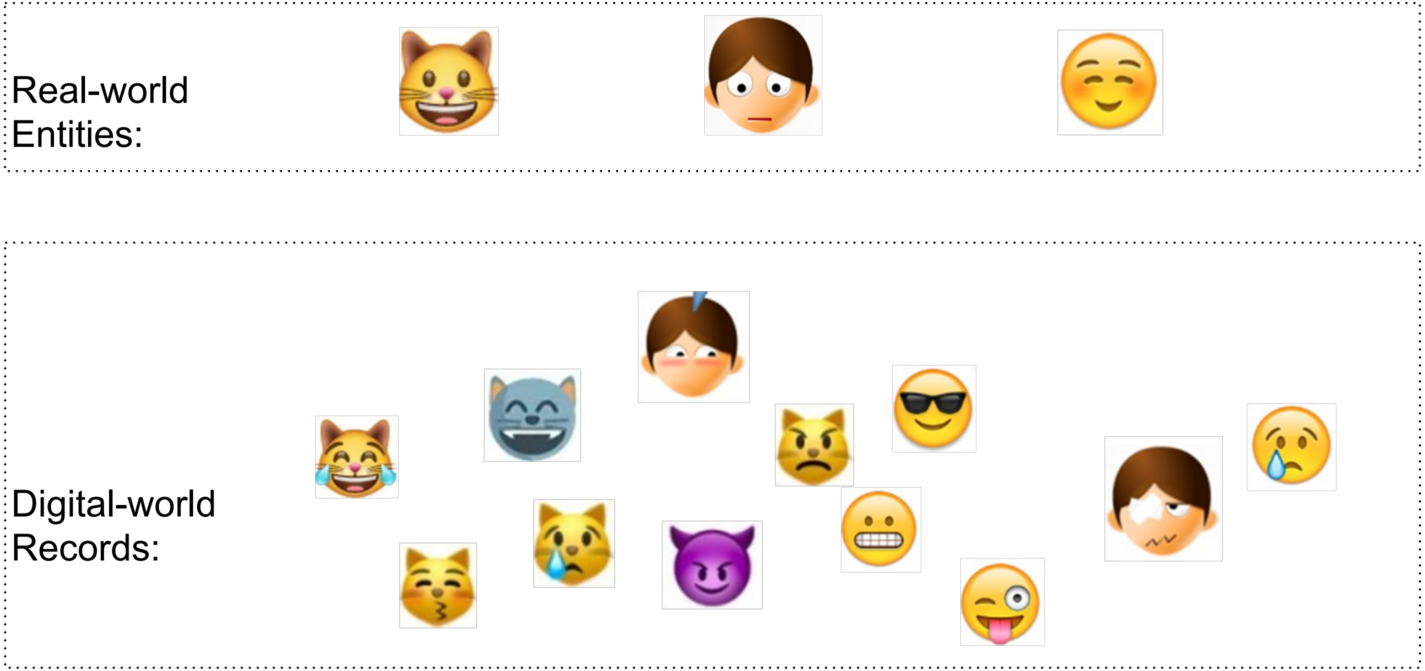
The Best of Both Worlds: Combining Hand-Tuned and Word-Embedding-Based Similarity Measures for Entity Resolution

Xiao Chen, Gabriel Campero Durand, Roman Zoun, David Broneske, Yang Li, Gunter Saake
xiao.chen@ovgu.de
Otto-von-Guericke-University of Magdeburg

BTW’19, Rostock, March 7th, 2019

Entity Resolution (ER)

- Real world vs. Digital world

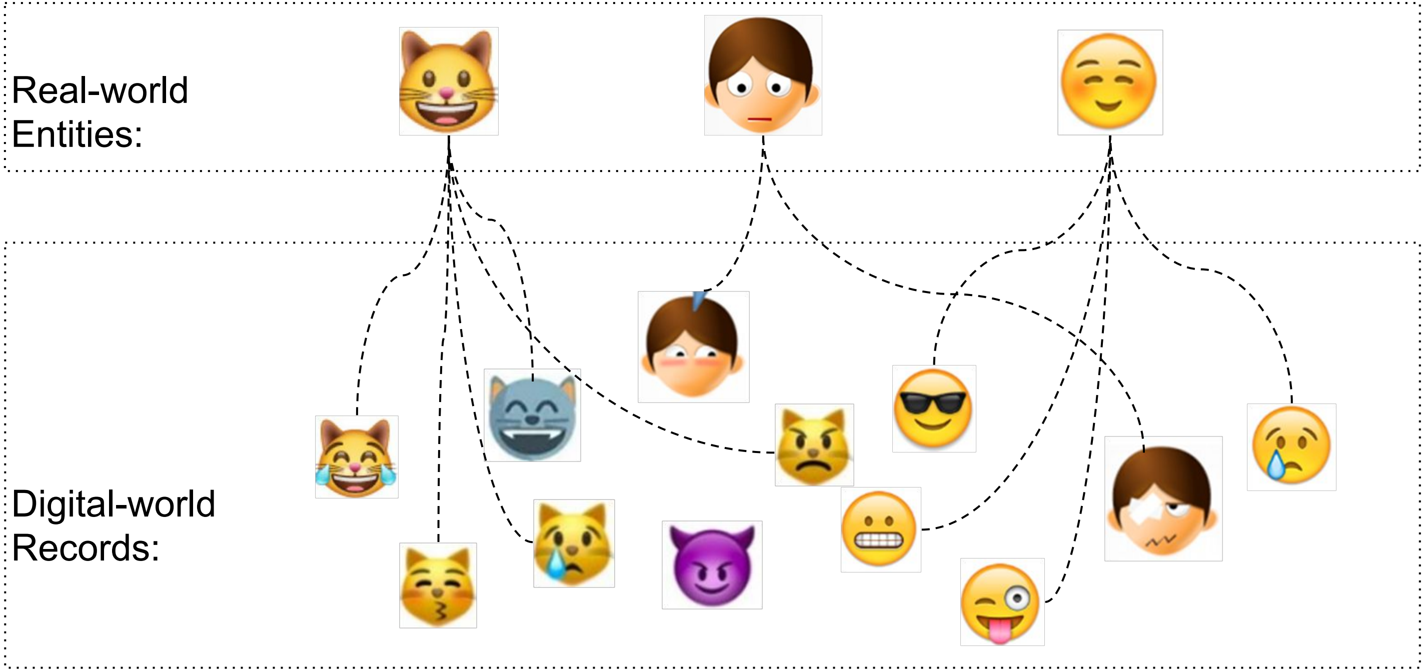


Real-world Entities:

Digital-world Records:

Entity Resolution (ER)

- Real world vs. Digital world
- Definition: Identifying records that refer to the same entity



Entity Resolution (ER)

- Real world vs. Digital world
- Definition: Identifying records that refer to the same entity

	Given-name	Surname	city	Postcode	Age	Phone-number	Sex
Hospital	starab	Kuaririo	brisbane	1402	25	03 2867 8172	f
Citizen's office	sarah	Guarino	brisbane	1402	26	03 2897 8172	m

Entity Resolution (ER)

- Real world vs. Digital world
- Definition: Identifying records that refer to the same entity

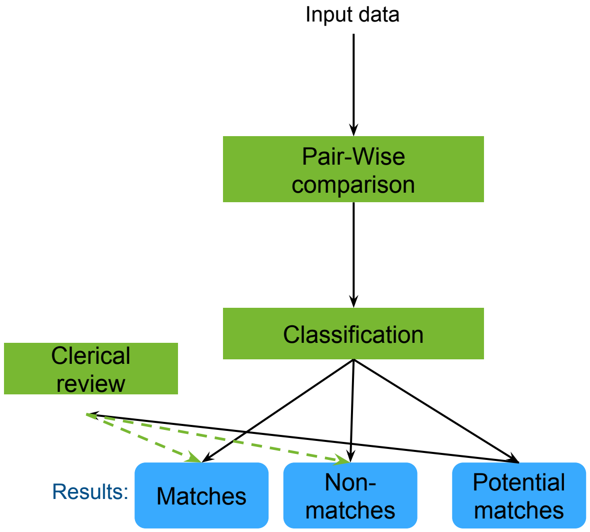
	Name	Description	Manufacturer	Price
Amazon	world book encyclopedia 2006	the world book encyclopedia 2006 is a truly student-friendly cd reference resource. it's been …	topics entertainment	19.99
Google	world book 2006	overview with over 87 years of experience and a global reputation for unsurpassed excellence world book 2006 is firmly established as the premier reference source for …	-	17.9

Entity Resolution (ER)

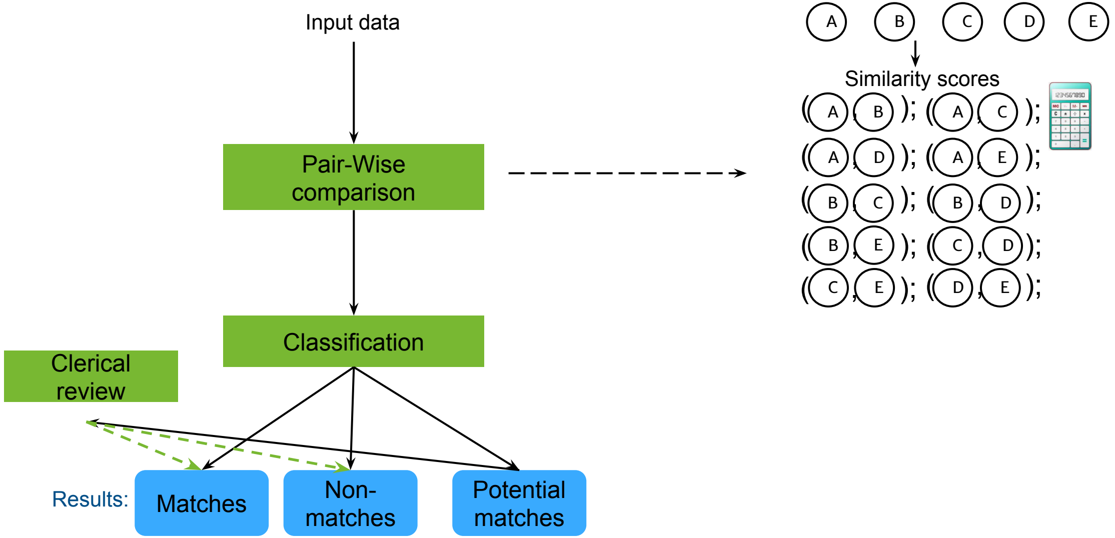
- Real world vs. Digital world
- Definition: Identifying records that refer to the same entity

	ID	Titel	Author	Venue	Year
DBLP	conf/sigmod/GrossmanHQ95	PTool: A Light Weight Persistent Object Manager	David Hanley, Robert L. Grossman, Xiao Qin	SIGMOD Conference	1995
ACM	223901	PTool: a light weight persistent object manager	R. L. Grossman, D. Hanley, X. Qin	International Conference on Management of Data	1995

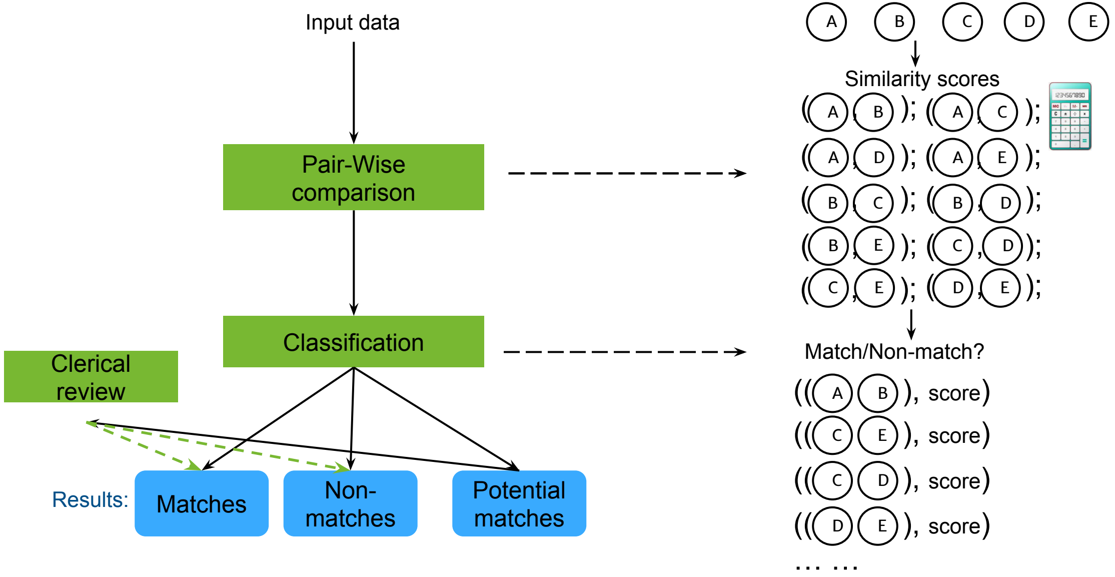
Basic Steps of Pair-Wise ER



Basic Steps of Pair-Wise ER



Basic Steps of Pair-Wise ER



Three Groups of Attributes

Persons:

Given-name	Surname	city	Postcode	Age	Phone-number	Sex
starab	Kuaririo	brisbane	1402	25	03 2867 8172	f
sarah	Guarino	brisbane	1402	26	03 2897 8172	m

DBLP-ACM bibliography data:

Titel	Author	Venue	Year
PTool: A Light Weight Persistent Object Manager	David Hanley, Robert L. Grossman, Xiao Qin	SIGMOD Conference	1995
PTool: a light weight persistent object manager	R. L. Grossman, D. Hanley, X. Qin	International Conference on Management of Data	1995

Amazon-Google product data:

Name	Description	Manufacturer	Price
world book encyclopedia 2006	the world book encyclopedia 2006 is a truly student-friendly cd reference resource. it's been …	topics entertainment	19.99
world book 2006	overview with over 87 years of experience and a global reputation for unsurpassed excellence world book 2006 is firmly established as the premier reference source for students parents teachers and librarians...	-	17.9

Three Groups of Attributes

- Numerical attributes (NA):

Persons:

Given-name	Surname	city	Postcode	**Age**	Phone-number	Sex
starab	Kuaririo	brisbane	1402	**25**	03 2867 8172	f
sarah	Guarino	brisbane	1402	**26**	03 2897 8172	m

DBLP-ACM bibliography data:

Titel	Author	Venue	Year
PTool: A Light Weight Persistent Object Manager	David Hanley, Robert L. Grossman, Xiao Qin	SIGMOD Conference	1995
PTool: a light weight persistent object manager	R. L. Grossman, D. Hanley, X. Qin	International Conference on Management of Data	1995

Amazon-Google product data:

Name	Description	Manufacturer	**Price**
world book encyclopedia 2006	the world book encyclopedia 2006 is a truly student-friendly cd reference resource. it's been …	topics entertainment	**19.99**
world book 2006	overview with over 87 years of experience and a global reputation for unsurpassed excellence world book 2006 is firmly established as the premier reference source for students parents teachers and librarians…		**17.9**

Three Groups of Attributes

- Numerical attributes (NA):
  - Don’t include numerical strings

Persons:

Given-name	Surname	city	Postcode	Age	Phone-number	Sex
starab	Kuaririo	brisbane	1402	25	03 2867 8172	f
sarah	Guarino	brisbane	1402	26	03 2897 8172	m

DBLP-ACM bibliography data:

Titel	Author	Venue	Year
PTool: A Light Weight Persistent Object Manager	David Hanley, Robert L. Grossman, Xiao Qin	SIGMOD Conference	1995
PTool: a light weight persistent object manager	R. L. Grossman, D. Hanley, X. Qin	International Conference on Management of Data	1995

Amazon-Google product data:

Name	Description	Manufacturer	Price
world book encyclopedia 2006	the world book encyclopedia 2006 is a truly student-friendly cd reference resource. it's been …	topics entertainment	19.99
world book 2006	overview with over 87 years of experience and a global reputation for unsurpassed excellence world book 2006 is firmly established as the premier reference source for students parents teachers and librarians…		17.9

Three Groups of Attributes

- Numerical attributes (NA):
- Non-semantically related attributes (NRA):
  - Often relatively short strings (including numerical strings)
  - Without semantics
  - Possible reasons: typos, formats

Persons:

Given-name	Surname	city	Postcode	Age	Phone-number	Sex
starab	Kuaririo	brisbane	1402	25	03 2867 8172	f
sarah	Guarino	brisbane	1402	26	03 2897 8172	m

DBLP-ACM bibliography data:

Titel	Author	Venue	Year
PTool: A Light Weight Persistent Object Manager	David Hanley, Robert L. Grossman, Xiao Qin	SIGMOD Conference	1995
PTool: a light weight persistent object manager	R. L. Grossman, D. Hanley, X. Qin	International Conference on Management of Data	1995

Amazon-Google product data:

Name	Description	Manufacturer	Price
world book encyclopedia 2006	the world book encyclopedia 2006 is a truly student-friendly cd reference resource. it's been …	topics entertainment	19.99
world book 2006	overview with over 87 years of experience and a global reputation for unsurpassed excellence world book 2006 is firmly established as the premier reference source for students parents teachers and librarians...		17.9

Three Groups of Attributes

- Numerical attributes (NA):
- Non-semantically related attributes (NRA):
  - ○ Often relatively short strings (including numerical strings)
  - ○ Without semantics
  - ○ Possible reasons: typos, formats
- Semantically related attributes (SRA):
  - ○ Often relatively long strings or sentences
  - ○ With semantics
  - ○ Possible reasons: different expressions, different names

Persons:

Given-name	Surname	city	Postcode	Age	Phone-number	Sex
starab	Kuaririo	brisbane	1402	25	03 2867 8172	f
sarah	Guarino	brisbane	1402	26	03 2897 8172	m

DBLP-ACM bibliography data:

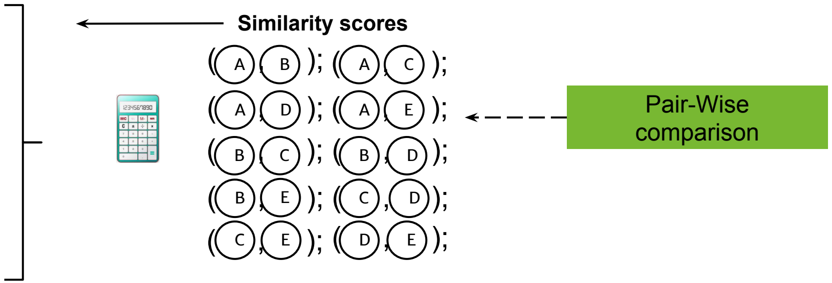
Titel	Author	Venue	Year
PTool: A Light Weight Persistent Object Manager	David Hanley, Robert L. Grossman, Xiao Qin	SIGMOD Conference	1995
PTool: a light weight persistent object manager	R. L. Grossman, D. Hanley, X. Qin	International Conference on Management of Data	1995

Amazon-Google product data:

Name	Description	Manufacturer	Price
world book encyclopedia 2006	the world book encyclopedia 2006 is a truly student-friendly cd reference resource. it's been …	topics entertainment	19.99
world book 2006	overview with over 87 years of experience and a global reputation for unsurpassed excellence world book 2006 is firmly established as the premier reference source for students parents teachers and librarians...		17.9

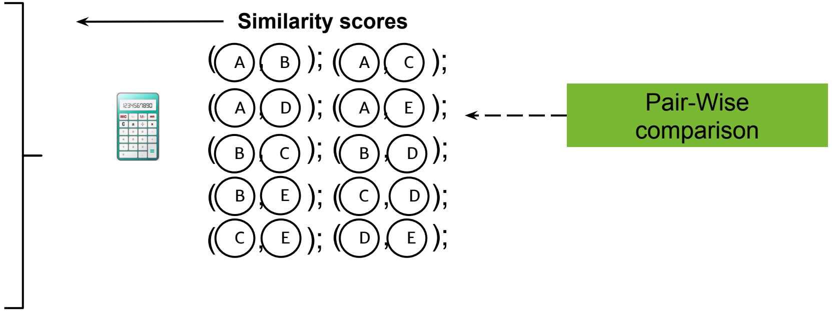
Approaches to Calculate Similarities

- Traditional approaches:
  - ○ Syntactical-based
  - ○ Without considering semantics
  - ○ Correct selection of similarity measures by domain experts



Approaches to Calculate Similarities

- Traditional approaches:
  - Syntactical-based
  - Without considering semantics
  - Correct selection of similarity measures by domain experts
    - ➢ Limited accuracy for SRAs



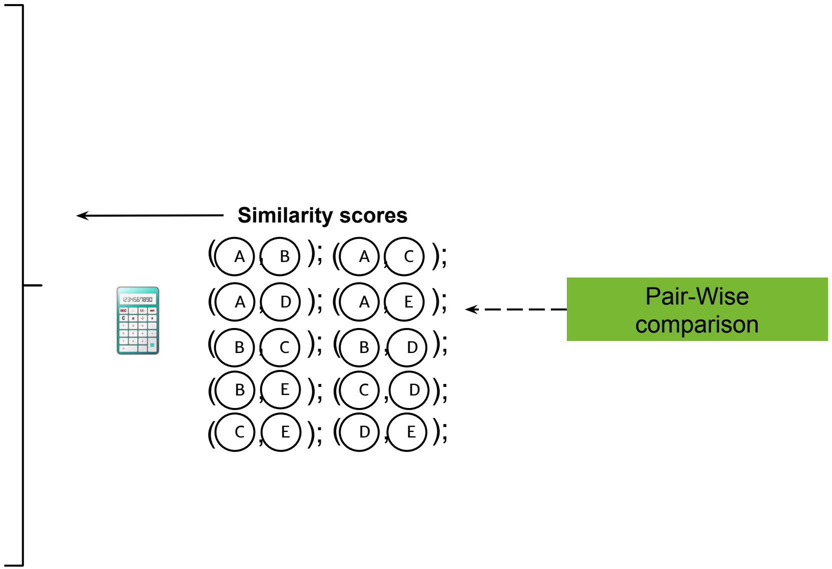
Similarity scores

(A, B); (A, C);
(A, D); (A, E);
(B, C); (B, D);
(B, E); (C, D);
(C, E); (D, E);

Pair-Wise comparison

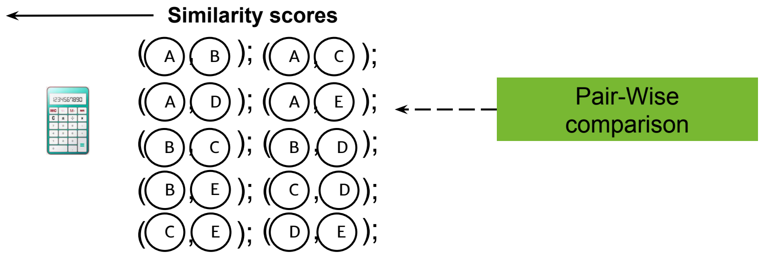
Approaches to Calculate Similarities

- Traditional approaches:
  - ○ Syntactical-based
  - ○ Without considering semantics
  - ○ Correct selection of similarity measures by domain experts
    - ➢ Limited accuracy for SRAs
- Recently:
  - ○ Word embedding based
  - ○ Considering semantics
  - ○ Applicable for all kinds of data



Approaches to Calculate Similarities

- Traditional approaches:
  - Syntactical-based
  - Without considering semantics
  - Correct selection of similarity measures by domain experts
    - ➢ Limited accuracy for SRAs
- Recently:
  - Word embedding based
  - Considering semantics
  - Applicable for all kinds of data
    - ➢ Negative effects on efficiency
    - ➢ Possible low accuracy for NAs and NRAs



Problems Using A Single Approach

- No one-fit-all solution
- One dataset contains more than one type of attributes:
  - Non-semantically related attributes (NRA)
  - Semantically related attributes (SRA)
  - Numerical attributes (NA)

Persons:

Given-name	Surname	city	Postcode	Age	Phone-number	Sex
starab	Kuaririo	brisbane	1402	25	03 2867 8172	f
sarah	Guarino	brisbane	1402	26	03 2897 8172	m

DBLP-ACM bibliography data:

Titel	Author	Venue	Year
PTool: A Light Weight Persistent Object Manager	David Hanley, Robert L. Grossman, Xiao Qin	SIGMOD Conference	1995
PTool: a light weight persistent object manager	R. L. Grossman, D. Hanley, X. Qin	International Conference on Management of Data	1995

Amazon-Google product data:

Name	Description	Manufacturer	Price
world book encyclopedia 2006	the world book encyclopedia 2006 is a truly student-friendly cd reference resource. it's been …	topics entertainment	19.99
world book 2006	overview with over 87 years of experience and a global reputation for unsurpassed excellence world book 2006 is firmly established as the premier reference source for students parents teachers and librarians...	-	17.9

Problems Using A Single Approach

- No one-fit-all solution
- One dataset contains more than one type of attributes:
  - Non-semantically related attributes (NRA)
  - Semantically related attributes (SRA)
  - Numerical attributes (NA)

➢ **Hybrid approach to calculate similarity scores**

Persons:

Given-name	Surname	city	Postcode	Age	Phone-number	Sex
starab	Kuaririo	brisbane	1402	25	03 2867 8172	f
sarah	Guarino	brisbane	1402	26	03 2897 8172	m

DBLP-ACM bibliography data:

Titel	Author	Venue	Year
PTool: A Light Weight Persistent Object Manager	David Hanley, Robert L. Grossman, Xiao Qin	SIGMOD Conference	1995
PTool: a light weight persistent object manager	R. L. Grossman, D. Hanley, X. Qin	International Conference on Management of Data	1995

Amazon-Google product data:

Name	Description	Manufacturer	Price
world book encyclopedia 2006	the world book encyclopedia 2006 is a truly student-friendly cd reference resource. it's been …	topics entertainment	19.99
world book 2006	overview with over 87 years of experience and a global reputation for unsurpassed excellence world book 2006 is firmly established as the premier reference source for students parents teachers and librarians...	-	17.9

Hybrid Approach

- Non-semantically related attributes (NRA):
  - Relatively short strings (including numerical strings)
  - Without semantics
- Numerical attributes (NA):

Traditional approaches

- Syntactical-based
- Without considering semantics
- Choosing suitable functions:

$$attrSim(r_1.attr, r_2.attr) = \begin{cases} Euclidean(r_1.attr, r_2.attr), & attr \in NA; \\ Jaro\_Winkler(r_1.attr, r_2.attr), & attr \in NRA. \end{cases}$$

Hybrid Approach

- Non-semantically related attributes (NRA):
  - Relatively short strings (including numerical strings)
  - Without semantics
- Numerical attributes (NA):

Traditional approaches

- Syntactical-based
- Without considering semantics
- Choosing suitable functions:

$$attrSim(r_1.attr, r_2.attr) = \begin{cases} Euclidean(r_1.attr, r_2.attr), & attr \in NA; \\ Jaro\_Winkler(r_1.attr, r_2.attr), & attr \in NRA. \end{cases}$$

- Semantically related attributes (SRA):
  - Relatively long strings
  - With semantics

Word embedding based

- Considering semantics
- Cosine similarity on transformed vectors

Word Embedding Approach for SRAs

- Vector for one word:
  - FastText model

Word Embedding Approach for SRAs

- Vector for one word:
  - FastText model
- Vector for one attribute:
  - $$\vec{attr} = \frac{\sum_{i=1}^{n} \vec{w_i}}{n}$$

Word Embedding Approach for SRAs

- Vector for one word:
  - FastText model
- Vector for one attribute:
  - $\vec{attr} = \frac{\sum_{i=1}^{n} \vec{w_i}}{n}$
- Similarity scores calculated on each attribute vector:
  - $attrSim(r_1.attr, r_2.attr) = cosine(r_1.\vec{attr}, r_2.\vec{attr}), \quad attr \in SRA.$

Evaluation: Setup

- Three datasets:

Datasets	#Pairs (DS1 & DS2)	#Matches
Persons	551250 (1050 & 1050)	96
DBLP - ACM	6001104 (2616 & 2294)	2224
Amazon - Google	4400264 (1364 & 3226)	1300

Persons:

Given-name	Surname	city	Postcode	Age	Phone-number	Sex
starab	Kuaririo	brisbane	1402	25	03 2867 8172	f
sarah	Guarino	brisbane	1402	26	03 2897 8172	m

DBLP-ACM bibliography data:

Titel	Author	Venue	Year
PTool: A Light Weight Persistent Object Manager	David Hanley, Robert L. Grossman, Xiao Qin	SIGMOD Conference	1995
PTool: a light weight persistent object manager	R. L. Grossman, D. Hanley, X. Qin	International Conference on Management of Data	1995

Amazon-Google product data:

Name	Description	Manufacturer	Price
world book encyclopedia 2006	the world book encyclopedia 2006 is a truly student-friendly cd reference resource. it's been …	topics entertainment	19.99
world book 2006	overview with over 87 years of experience and a global reputation for unsurpassed excellence world book 2006 is firmly established as the premier reference source for students parents teachers and librarians...	-	17.9

Evaluation: Setup

- Three datasets:

Datasets	#Pairs (DS1 & DS2)	#Matches	#SRAs	#NRAs	#NAs
Persons	551250 (1050 & 1050)	96	2	6	5
DBLP - ACM	6001104 (2616 & 2294)	2224	2	2	0
Amazon - Google	4400264 (1364 & 3226)	1300	3	0	1

Evaluation: Setup

- Approaches for similarity calculations:
  - Traditional similarity functions only:
    - Jaro-Winkler for SRAs and NRAs
    - Euclidean distance for NAs
  - Word embedding and cosine similarity based method only:
    - Word embedding + cosine similarity for all SRAs, NRAs and NAs
  - Hybrid:
    - Jaro-Winkler for NRAs
    - Euclidean distance for NAs
    - Word embedding + cosine similarity for SRAs

Evaluation: Setup

- Classification approach: learning-based classification
  - XGBoost
  - Random forest
  - K-Nearest neighbor

Evaluation: Setup

- Classification approach: learning-based classification
  - XGBoost
  - Random forest
  - K-Nearest neighbor

- Training & test data:
  - Took all pairs of cartesian product;
  - For training, 66% of matches & 66% of non-matches;
  - For testing, remaining 34% of both.

Evaluation: Results

- Persons:
  - Best: word embedding
  - KNN F-measures

Combinations		XGBoost	RF	KNN
Persons	Traditional	100	100	88.46
	WordEmbedding	**100**	**100**	**100**
	Hybrid	100	100	58.54

Evaluation: Results

- Persons:
  - Best: word embedding
  - KNN F-measures
- DBLP - ACM bibliography:
  - Best: traditional approach
  - “Title” should belong to NRA

Combinations		XGBoost	RF	KNN
Persons	Traditional	100	100	88.46
	WordEmbedding	100	100	100
	Hybrid	100	100	58.54
DBLP - ACM	Traditional	**97.04**	**97.7**	**95.17**
	WordEmbedding	92.56	94.82	93.94
	Hybrid	93.69	94.28	89.31

Evaluation: Results

- Persons:
  - Best: word embedding
  - KNN F-measures
- DBLP - ACM bibliography:
  - Best: traditional approach
  - “Title” should belong to NRA
- Amazon - Google product:
  - Word-Embedding outperforms traditional for RF and KNN, is comparable for XGBoost
  - Hybrid approach is the best for XGBoost and RF

Combinations		XGBoost	RF	KNN
Persons	Traditional	100	100	88.46
	WordEmbedding	100	100	100
	Hybrid	100	100	58.54
DBLP - ACM	Traditional	97.04	97.7	95.17
	WordEmbedding	92.56	94.82	93.94
	Hybrid	93.69	94.28	89.31
Amazon - Google	Traditional	20.19	25.35	21.11
	WordEmbedding	19.10	31.09	24.1
	Hybrid	**29.72**	**38.32**	19.78

Evaluation: Results

- A true matching example of a product pair:

Amazon:

train sim modeler design studio, with train sim modeler you can create 3d traincars boxcars and engines along with your own custom scenery! create train station stores hills and trees and more scenery set up a virtual cab so you can see from the train driver's view you'll have your own personal railroad cars running the rails in no time!,abacus,39.99

Google:

train sim modeler, microsoft train simulator brings the most realistic virtual train experience to the pc. already ms train simulator is the number one selling simulator in europe. and by all indications microsoft train simulator (ts) is a bestseller since it was ..., ,29.84

	Name	**Description**	**Manufacturer**	**Price**
Traditional	0.6611724	0.72039728	0.0	0.99997712
WordEmbedding	0.8569186	0.87175614	0.0	-0.03565185
Hybrid	0.8569186	0.87175614	0.0	0.99997712

Evaluation: Results

- Lower than published results

Combinations		**XGBoost**	**RF**	**KNN**
Persons	Traditional	100	100	88.46
	WordEmbedding	100	100	100
	Hybrid	100	100	58.54
DBLP - ACM	Traditional	97.04	97.7	95.17
	WordEmbedding	92.56	94.82	93.94
	Hybrid	93.69	94.28	89.31
Amazon - Google	Traditional	**20.19**	**25.35**	**21.11**
	WordEmbedding	**19.10**	**31.09**	**24.1**
	Hybrid	**29.72**	**38.32**	**19.78**

Evaluation: Results

- Word embedding:
  - SRAs: predominantly better
  - NRAs: comparable or worse
  - NAs: not recommended
- Hybrid approach:
  - Is able to provide better accuracy for data including different types of attributes
- Classifier choices

Combinations		XGBoost	RF	KNN
Persons	Traditional	100	100	88.46
	WordEmbedding	100	100	100
	Hybrid	100	100	58.54
DBLP - ACM	Traditional	97.04	97.7	95.17
	WordEmbedding	92.56	94.82	93.94
	Hybrid	93.69	94.28	89.31
Amazon - Google	Traditional	20.19	25.35	21.11
	WordEmbedding	19.10	31.09	24.1
	Hybrid	29.72	38.32	19.78

Conclusion

- Three groups of attributes:
  - SRAs, NRAs and NAs
- Hybrid similarity calculations:
  - SRAs: word embedding + cosine similarity
  - NRAs and NAs: traditional similarity functions
- Evaluation:
  - Word embedding performs predominantly better for SRAs, and worse for NAs;
  - Hybrid approach is useful to fix the similarity scores, which are wrongly calculated by word embedding for numerical attributes.

Future Work

- Evaluate the hybrid approach when using blocking or thresholding techniques
- Classification algorithms

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

INF FAKULTÄT FÜR INFORMATIK

DBSE

Thank you!

Xiao Chen, Gabriel Campero Durand, Roman Zoun, David Broneske, Yang Li, Gunter Saake
xiao.chen@ovgu.de
Otto-von-Guericke-University of Magdeburg

BTW’19, Rostock, March 7th, 2019

References

[1] Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606, 2016.

[2] Chen, T.; Guestrin, C.: Xgboost: A scalable tree boosting system. In: SIGKDD. ACM, pp. 785–794, 2016.

[3] Ebraheem, M.; Thirumuruganathan, S.; Joty, S. R.; Ouzzani, M.; Tang, N.: Distributed Representations of Tuples for Entity Resolution. PVLDB, 11(11):1454–1467, 2018.

[4] Kooli, N.; Allesiardo, R.; Pigneul, E.: Deep Learning Based Approach for Entity Resolution in Databases. In: ACIIDS. Springer, pp. 3–12, 2018.

[5] Köpcke, H.; Rahm, E.: Training selection for tuning entity matching. In: QDB/MUD. pp. 3–12, 2008.

[6] Mikolov, T.; Chen, K.; Corrado, G.; Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[7] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. Curran Associates, 2013.

[8] Mudgal, S.; Li, H.; Rekatsinas, T.; Doan, A.; Park, Y.; Krishnan, G.; Deep, R.; Arcaute, E.; Raghavendra, V.: Deep Learning for Entity Matching: A Design Space Exploration. In: SIGMOD. ACM, pp. 19–34, 2018.

[9] Pennington, J.; R. Socher, Riand Manning, Christopher: Glove: Global vectors for word representation. In: EMNLP. pp. 1532–1543, 2014.