Interlinking: Performance Assessment of User Evaluation vs. Supervised Learning Approaches

Mofeed Hassan, Jens Lehmann and Axel-Cyrille Ngonga Ngomo

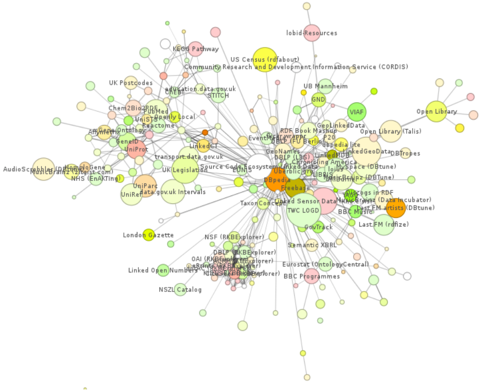
Agile Knowledge Engineering and Semantic Web
Department of Computer Science
University of Leipzig
Augustusplatz 10, 04109 Leipzig

{mounir,lehmann,ngonga}@informatik.uni-leipzig.de
WWW home page: http://limes.sf.net

May 17, 2015

Why Link Discovery?

1. Fourth Linked Data principle
2. Links are central for
   - Cross-ontology QA
   - Data Integration
   - Reasoning
   - Federated Queries
   - ...
3. Linked Data on the Web:
   - 10+ thousand datasets
   - 89+ billion triples
   - ≈ 500+ million links



Why is it difficult?

Definition (Link Discovery)

- Given sets S and T of resources and relation $\mathcal{R}$
- Task: Find $M = \{(s,t) \in S \times T : \mathcal{R}(s,t)\}$
- Common approaches:
  - Find $M' = \{(s,t) \in S \times T : \sigma(s,t) \geq \theta\}$
  - Find $M' = \{(s,t) \in S \times T : \delta(s,t) \leq \theta\}$

1. **Time complexity**
   - Large number of triples
   - Quadratic a-priori runtime
   - 69 days for mapping cities from DBpedia to Geonames (1ms per comparison)
   - Decades for linking DBpedia and LGD ...

Why is it difficult?

Definition (Link Discovery)

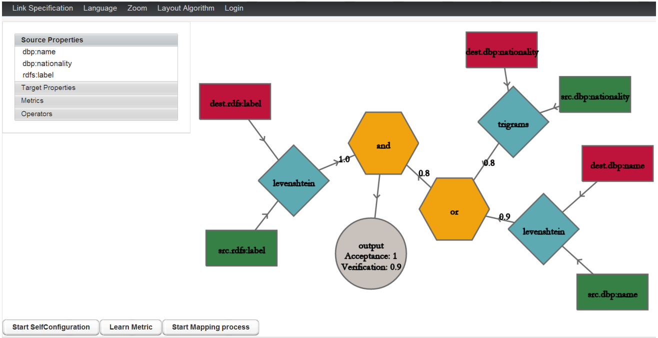
- Given sets S and T of resources and relation $\mathcal{R}$
- Task: Find $M = \{(s, t) \in S \times T : \mathcal{R}(s, t)\}$
- Common approaches:
  - Find $M' = \{(s, t) \in S \times T : \sigma(s, t) \geq \theta\}$
  - Find $M' = \{(s, t) \in S \times T : \delta(s, t) \leq \theta\}$

1. **Time complexity**
   - Large number of triples
   - Quadratic a-priori runtime
   - 69 days for mapping cities from DBpedia to Geonames (1ms per comparison)
   - Decades for linking DBpedia and LGD ...

Why is it difficult?

2 Complexity of specifications

- Combination of several attributes required for high precision
- **Adequate atomic similarity functions difficult to detect**
- Tedious discovery of most adequate mapping



Introduction

- Interlinking tools LIMES, SILK, RDFAI,...
- Interlinking tools differ in many factors such as:
  1. Automation and user involvement
  2. Domain dependency
  3. Matching techniques
- Manual links validation as a user involvement:
  1. Benchmarks
  2. Active learning positive and negative examples

Introduction

- Commonly used
  - String distance/similarity measures
    - **Edit distance**
    - Q-Gram similarity
    - Jaro-Winkler
    - ...
  - Metrics
    - Minkowski distance
    - Orthodromic distance
    - Symmetric Hausdorff distance
    - ...

Idea

- Learning distance/similarity measures from data can lead to better accuracy while linking.

Introduction

- Commonly used
  - String distance/similarity measures
    - **Edit distance**
    - Q-Gram similarity
    - Jaro-Winkler
    - ...
  - Metrics
    - Minkowski distance
    - Orthodromic distance
    - Symmetric Hausdorff distance
    - ...

Idea

- Learning distance/similarity measures from data can lead to better accuracy while linking.

Motivation/1

Problem

- Edit distance does not differentiate between different types of edits.

Source labels	Target labels
Generalised epidermolysis	Generalized epidermolysis
Diabetes I	Diabetes I
Diabetes II	Diabetes II

Motivation/1

Problem

- Edit distance does not differentiate between different types of edits.

Source labels	Target labels
Generali<u>s</u>ed epidermolysis	Generali<u>z</u>ed epidermolysis
Diabetes <u>I</u>	Diabetes I
Diabetes II	Diabetes <u>II</u>

Motivation/2

- Choosing $\theta \in [0, 1)$

	%
F-Score	**80.0**
Precision	100.0
Recall	66.7

- Choosing $\theta \in [1, 2)$

	%
F-Score	**75.0**
Precision	60.0
Recall	100.0

Solution: Weighted edit distance

- Assign weight to each operation: substitution, insertion, deletion.

Motivation/2

- Choosing $\theta \in [0, 1)$

	%
F-Score	**80.0**
Precision	100.0
Recall	66.7

- Choosing $\theta \in [1, 2)$

	%
F-Score	**75.0**
Precision	60.0
Recall	100.0

Solution: Weighted edit distance

- Assign weight to each operation: substitution, insertion, deletion.

Motivation/3

Cost matrix

- Costs are arranged in a quadratic matrix M
- Cell $m_{i,j}$ contains the cost of transforming character associated to row i into character associated with column j
- Characters are from an alphabet $\{`A`, \ldots, `Z`, `a`, \ldots, `z`, `0`, \ldots, `9`, `\epsilon`\}$
- Main diagonal values are zeros

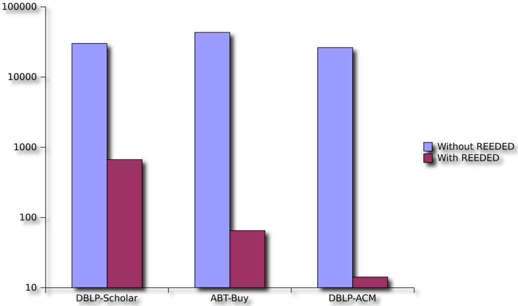
$$
\begin{array}{c|cccccc|}
 & \mathbf{A} & \mathbf{B} & \mathbf{C} & \mathbf{D} & \cdots & \boldsymbol{\varepsilon} \\
\hline
\mathbf{A} & 0 & 1 & 1 & 1 & \cdots & 1 \\
\mathbf{B} & 1 & 0 & 1 & 1 & \cdots & 1 \\
\mathbf{C} & 1 & 1 & 0 & 1 & \cdots & 1 \\
\mathbf{D} & 1 & 1 & 1 & 0 & \cdots & 1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\boldsymbol{\varepsilon} & 1 & 1 & 1 & 1 & \cdots & 0 \\
\hline
\end{array}
$$

Motivation/4

- Pros
  - Can differentiate between edit operations.
  - Better F-measure in some cases.
- Cons
  - No dedicated scalable algorithm for weighted edit distances
  - Difficult to use for link discovery.

Motivation/5

	DBLP–Scholar	ABT–Buy	DBLP–ACM
F-measure (%)	87.85	0.60	97.92
Without REEDED (s)	30,096	43,236	26,316
With REEDED (s)	668.62	65.21	14.24



Extension of existing algorithms

Idea

- $edit(x, y) = \theta \rightarrow$ Need θ operations to transform x into y
- $\delta(x, y) \geq \theta \cdot \min_{i \neq j} m_{ij}$

Extension

1. Run existing algorithm with threshold $\frac{\theta}{\min_{i \neq j} m_{ij}}$
2. Filter results by using $\delta(x, y) \geq \theta$

Problem

Does not scale.

Extension of existing algorithms

Idea

- $edit(x, y) = \theta \rightarrow$ Need θ operations to transform x into y
- $\delta(x, y) \geq \theta \cdot \min_{i \neq j} m_{ij}$

Extension

1. Run existing algorithm with threshold $\frac{\theta}{\min_{i \neq j} m_{ij}}$
2. Filter results by using $\delta(x, y) \geq \theta$

Problem

Does not scale.

Extension of existing algorithms

Idea

- $edit(x, y) = \theta \rightarrow$ Need θ operations to transform x into y
- $\delta(x, y) \geq \theta \cdot \min_{i \neq j} m_{ij}$

Extension

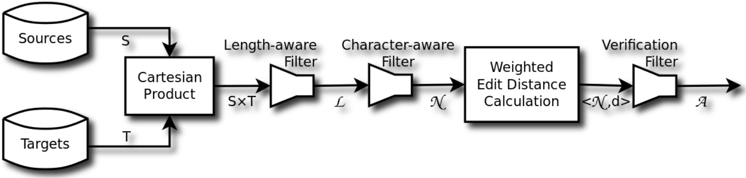
1. Run existing algorithm with threshold $\frac{\theta}{\min_{i \neq j} m_{ij}}$
2. Filter results by using $\delta(x, y) \geq \theta$

Problem

Does not scale.

REEDED

- Series of filters.
- Both **complete** and **correct**.



Length-Aware Filter

- *Input*: a pair $(s, t) \in S \times T$ and a threshold θ
- *Output*: the pair itself or null

Insight

Given two strings s and t with lengths $|s|$ resp. $|t|$, we need at least $||s| - |t||$ edit operations to transform s into t.

Examples

A. $\langle s, t, \theta \rangle = \langle$ "*realize*", "*realise*", $1 \rangle$
$||s| - |t|| = 0, \quad \Rightarrow$ *pass*
B. $\langle s, t, \theta \rangle = \langle$ "*realize*", "*real*", $1 \rangle$
$||s| - |t|| = 3, \quad \Rightarrow$ *discard*

Character-Aware Filter

- *Input*: a pair $(s, t) \in \mathcal{L}$ and a threshold θ
- *Output*: the pair itself or null

Insight

Given two strings s and t, if $|C|$ is the number of characters that do not belong to both strings, we need at least $\frac{|C|}{2}$ operations to transform s into t.

Examples

A. $\langle s, t, \theta \rangle = \langle$ "*realize*", "*realise*", $1 \rangle$

$C = \{s, z\}$, $\lfloor \frac{|C|}{2} \rfloor \cdot \min\limits_{i \neq j}(m_{ij}) = 0.5$, $\Rightarrow$ *pass*

B. $\langle s, t, \theta \rangle = \langle$ "*realize*", "*concept*", $1 \rangle$

$C = \{r, c, a, l, i, z, o, n, p, t\}$, $\lfloor \frac{|C|}{2} \rfloor \cdot \min\limits_{i \neq j}(m_{ij}) > 1$, $\Rightarrow$ *discard*

Verification Filter

- *Input*: a pair $(s, t) \in \mathcal{C}$ and a threshold θ
- *Output*: the pair itself or null

Insight

Definition of Weighted Edit Distance. Two strings s and t are similar iff the sum of the operation costs to transform s into t is less than or equal to θ.

Examples

A. $\langle s, t, \theta \rangle = \langle$"*realize*", "*realise*", $1\rangle$
$\delta(s, t) = m_{z,s} = 0.6, \qquad \Rightarrow$ *pass*

Experimental Setup/1

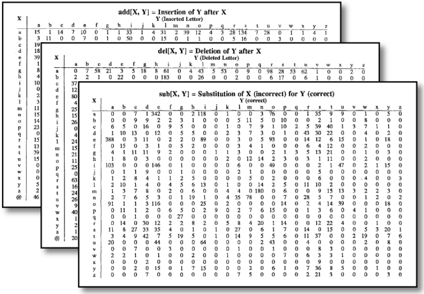
Datasets

dataset.property	domain	# of pairs	avg length
DBLP.title	bibliographic	6,843,456	56.359
ACM.authors	bibliographic	5,262,436	46.619
GoogleProducts.name	e-commerce	10,407,076	57.024
ABT.description	e-commerce	1,168,561	248.183

Experimental Setup/2

Weight configuration

Given an edit operation, the higher the probability of error, the lower its weight.

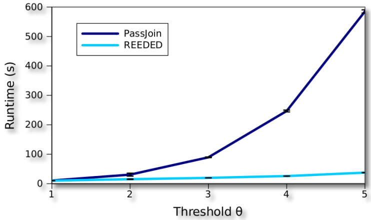


1. Load typographical error frequencies
2. For insertion and deletion, calculate total frequency for each character
3. Normalize values on max frequency

Evaluation/1

DBLP.title — bibliographic domain — 6,843,456 pairs

	PassJoin*		REEDED	
θ	average	st.dev.	average	st.dev.
1	10.75	± 0.92	10.38	± 0.35
2	30.74	± 5.00	15.27	± 0.76
3	89.60	± 1.16	19.84	± 0.14
4	246.93	± 3.08	25.91	± 0.29
5	585.08	± 5.47	37.59	± 0.43

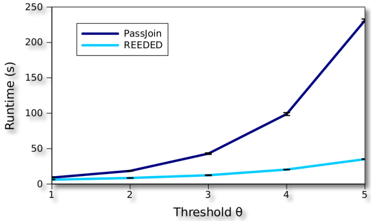


⋆ Extended to deal with weighted edit distances.

Evaluation/2

ACM.authors — bibliographic domain — 5,262,436 pairs

	PassJoin*		REEDED	
θ	average	st.dev.	average	st.dev.
1	9.07	$\pm$ 1.05	6.16	$\pm$ 0.07
2	18.53	$\pm$ 0.22	8.54	$\pm$ 0.29
3	42.97	$\pm$ 1.02	12.43	$\pm$ 0.47
4	98.86	$\pm$ 1.98	20.44	$\pm$ 0.27
5	231.11	$\pm$ 2.03	35.13	$\pm$ 0.35

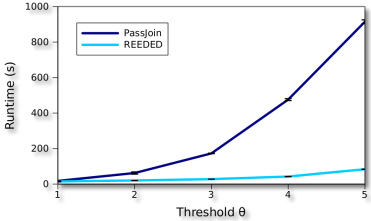


⋆ Extended to deal with weighted edit distances.

Evaluation/3

GoogleProducts.name — e-commerce domain — 10,407,076 pairs

	PassJoin*		REEDED	
θ	average	st.dev.	average	st.dev.
1	17.86	± 0.22	15.08	± 2.50
2	62.31	± 6.30	20.43	± 0.10
3	172.93	± 1.59	27.99	± 0.19
4	475.97	± 5.34	42.46	± 0.32
5	914.60	± 10.47	83.71	± 0.97

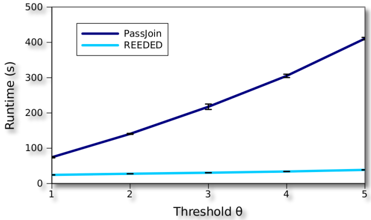


⋆ Extended to deal with weighted edit distances.

Evaluation/4

ABT.description — e-commerce domain — 1,168,561 pairs

	PassJoin*		REEDED	
θ	average	st.dev.	average	st.dev.
1	74.41	± 1.80	24.48	± 0.41
2	140.73	± 1.40	27.71	± 0.29
3	217.55	± 7.72	30.61	± 0.34
4	305.08	± 4.78	34.13	± 0.30
5	410.72	± 3.36	38.73	± 0.44



⋆ Extended to deal with weighted edit distances.

Effect of filters

GooglePr.name	$\theta = 1$	$\theta = 2$	$\theta = 3$	$\theta = 4$	$\theta = 5$
$\|S \times T\|$	10,407,076	10,407,076	10,407,076	10,407,076	10,407,076
$\|\mathcal{L}\|$	616,968	1,104,644	1,583,148	2,054,284	2,513,802
$\|\mathcal{N}\|$	4,196	4,720	9,278	38,728	153,402
$\|\mathcal{A}\|$	4,092	4,153	4,215	4,331	4,495
$RR(\%)$	99.96	99.95	99.91	99.63	95.53
ABT.description	$\theta = 1$	$\theta = 2$	$\theta = 3$	$\theta = 4$	$\theta = 5$
$\|S \times T\|$	1,168,561	1,168,561	1,168,561	1,168,561	1,168,561
$\|\mathcal{L}\|$	22,145	38,879	55,297	72,031	88,299
$\|\mathcal{N}\|$	1,131	1,193	1,247	1,319	1,457
$\|\mathcal{A}\|$	1,087	1,125	1,135	1,173	1,189
$RR(\%)$	99.90	99.90	99.89	99.88	99.87

Conclusion and Future Work

- Presented REEDED, a **time-efficient**, **correct** and **complete** LD approach for weighted edit distances
- Showed that REEDED scales better than simple extension of existing
- Future work includes:
  - Develop similar approach for weighted *n*-gram similarities.
  - Combine REEDED with specification learning approaches:
    - RAVEN, using Linear SVMs;
    - EAGLE, COALA using genetic programming.
  - Devise unsupervised learning approach for weights.

Thank you!
Questions?

Axel Ngonga
University of Leipzig
AKSW Research Group
Augustusplatz 10, Room P616
04109 Leipzig, Germany
ngonga@informatik.uni-leipzig.de