Efficient Bounded Jaro-Winkler Similarity Based Search

Jan Martin Keil

Heinz Nixdorf Chair for Distributed Information Systems
Institute for Computer Science
Friedrich Schiller University Jena, Germany

jan-martin.keil@uni-jena.de

18. BTW 2019 Uni Rostock 600 Jahre

7th of March 2019

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

FUSION

Use Cases

- Data
- Meta Data
- Search Queries

Research Data Management System

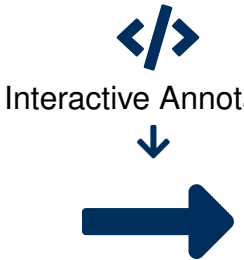
Use Cases

- Data
- Meta Data
- Search Queries

Research Data Management System

Use Cases

- Data
- Meta Data
- Search Queries



Interactive Annotation

Research Data Management System

- Spelling Mistakes
- Spelling Variants (e.g. *meter* and *metre*, species name suffix)

Use Cases

Approximate string matching for named entity identification

Jaro-Winkler Similarity

- Jaro-Winkler Similarity[1,2] is a similarity measure (not a metric) for short strings
- good general evaluation results[3]
- first characters emphasized
  - → spelling mistakes typically occur later[4]
  - → varying suffix tolerant

[1] Winkler 1990. “String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage”

[2] Winkler et al. 1994. strcmp95.c, Version 2 (original implementation)

[3] Cohen et al. 2003. “A Comparison of String Distance Metrics for Name-Matching Tasks”

[4] Pollock et al. 1983. “Collection and characterization of spelling errors in scientific and scholarly text”

Jaro Similarity

$$\text{Jaro}(s_1, s_2) = \begin{cases} \frac{1}{3} \times \left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m} \right) & : m > 0 \\ 0 & : \text{otherwise} \end{cases}$$

m: number of matching characters with max distance w

w: max distance of matching characters

t: number of transpositions

Jaro-Winkler Similarity

$$\text{JaroWinkler}(s_1, s_2) = \begin{cases} \text{Jaro}(s_1, s_2) + l \times p \times (1 - \text{Jaro}(s_1, s_2)) & : \text{Jaro}(s_1, s_2) \geq b_t \\ \text{Jaro}(s_1, s_2) & : \text{otherwise} \end{cases}$$

l: length of common prefix up to l_{limit}

l_{limit}: max length of common prefix $= 4$ [1]

p: prefix scale $= 0.1$ [1]

b_t: boost threshold $= 0.7$ [1]

[1] Winkler et al. 1994.

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$|s_1| =$

$|s_2| =$

$w =$

$m =$

$t =$

$\text{Jaro}(s_1, s_2) \approx$

$l =$

$\text{JaroWinkler}(s_1, s_2) \approx$

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$|s_1| = 8$

$|s_2| = 12$

$w =$

$m =$

$t =$

$\text{Jaro}(s_1, s_2) \approx$

$l =$

$\text{JaroWinkler}(s_1, s_2) \approx$

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$$|s_1| = 8$$

$$|s_2| = 12$$

$$w =$$

$$m =$$

$$t =$$

$$\text{Jaro}(s_1, s_2) \approx$$

$$l =$$

$$\text{JaroWinkler}(s_1, s_2) \approx$$

$$w = \frac{\max(|s_1|, |s_2|)}{2} - 1$$

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$$|s_1| = 8$$

$$|s_2| = 12$$

$$w = 5$$

$$m =$$

$$t =$$

$$\text{Jaro}(s_1, s_2) \approx$$

$$l =$$

$$\text{JaroWinkler}(s_1, s_2) \approx$$

$$w = \frac{\max(|s_1|, |s_2|)}{2} - 1$$

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$|s_1| = 8$

$|s_2| = 12$

$w = 5$

$m =$

$t =$

$\text{Jaro}(s_1, s_2) \approx$

$l =$

$\text{JaroWinkler}(s_1, s_2) \approx$

$m =$ equal chars not further apart than w

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$|s_1| = 8$

$|s_2| = 12$

$w = 5$

$m =$

$t =$

$\text{Jaro}(s_1, s_2) \approx$

$l =$

$\text{JaroWinkler}(s_1, s_2) \approx$

$m =$ equal chars not further apart than w

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$|s_1| = 8$

$|s_2| = 12$

$w = 5$

$m =$

$t =$

$\text{Jaro}(s_1, s_2) \approx$

$l =$

$\text{JaroWinkler}(s_1, s_2) \approx$

$m =$ equal chars not further apart than w

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$|s_1| = 8$

$|s_2| = 12$

$w = 5$

$m =$

$t =$

$\text{Jaro}(s_1, s_2) \approx$

$l =$

$\text{JaroWinkler}(s_1, s_2) \approx$

$m =$ equal chars not further apart than w

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$|s_1| = 8$

$|s_2| = 12$

$w = 5$

$m =$

$t =$

$\text{Jaro}(s_1, s_2) \approx$

$l =$

$\text{JaroWinkler}(s_1, s_2) \approx$

$m =$ equal chars not further apart than w

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$|s_1| = 8$

$|s_2| = 12$

$w = 5$

$m =$

$t =$

$\text{Jaro}(s_1, s_2) \approx$

$l =$

$\text{JaroWinkler}(s_1, s_2) \approx$

$m =$ equal chars not further apart than w

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$|s_1| = 8$

$|s_2| = 12$

$w = 5$

$m =$

$t =$

$\text{Jaro}(s_1, s_2) \approx$

$l =$

$\text{JaroWinkler}(s_1, s_2) \approx$

$m =$ equal chars not further apart than w

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$|s_1| = 8$

$|s_2| = 12$

$w = 5$

$m =$

$t =$

$\text{Jaro}(s_1, s_2) \approx$

$l =$

$\text{JaroWinkler}(s_1, s_2) \approx$

$m =$ equal chars not further apart than w

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$|s_1| = 8$

$|s_2| = 12$

$w = 5$

$m =$

$t =$

$\text{Jaro}(s_1, s_2) \approx$

$l =$

$\text{JaroWinkler}(s_1, s_2) \approx$

$m =$ equal chars not further apart than w

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$|s_1| = 8$

$|s_2| = 12$

$w = 5$

$m =$

$t =$

$\text{Jaro}(s_1, s_2) \approx$

$l =$

$\text{JaroWinkler}(s_1, s_2) \approx$

$m =$ equal chars not further apart than w

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$|s_1| = 8$

$|s_2| = 12$

$w = 5$

$m =$

$t =$

$\text{Jaro}(s_1, s_2) \approx$

$l =$

$\text{JaroWinkler}(s_1, s_2) \approx$

$m =$ equal chars not further apart than w

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$|s_1| = 8$

$|s_2| = 12$

$w = 5$

$m =$

$t =$

$\text{Jaro}(s_1, s_2) \approx$

$l =$

$\text{JaroWinkler}(s_1, s_2) \approx$

$m =$ equal chars not further apart than w

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$|s_1| = 8$

$|s_2| = 12$

$w = 5$

$m =$

$t =$

$\text{Jaro}(s_1, s_2) \approx$

$l =$

$\text{JaroWinkler}(s_1, s_2) \approx$

$m =$ equal chars not further apart than w

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$|s_1| = 8$

$|s_2| = 12$

$w = 5$

$m = 7$

$t =$

$\text{Jaro}(s_1, s_2) \approx$

$l =$

$\text{JaroWinkler}(s_1, s_2) \approx$

$m =$ equal chars not further apart than w

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$$|s_1| = 8$$

$$|s_2| = 12$$

$$w = 5$$

$$m = 7$$

$$t =$$

$$\text{Jaro}(s_1, s_2) \approx$$

$$l =$$

$$\text{JaroWinkler}(s_1, s_2) \approx$$

$$t = \frac{\text{unequal positions in matching chars}}{2}$$

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$$|s_1| = 8$$

$$|s_2| = 12$$

$$w = 5$$

$$m = 7$$

$$t =$$

$$\text{Jaro}(s_1, s_2) \approx$$

$$l =$$

$$\text{JaroWinkler}(s_1, s_2) \approx$$

$$t = \frac{\text{unequal positions in matching chars}}{2}$$

m_1	M	e	o	s	i	r	a
m_2	M	e	r	i	s	o	a

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$$|s_1| = 8$$

$$|s_2| = 12$$

$$w = 5$$

$$m = 7$$

$$t =$$

$$\text{Jaro}(s_1, s_2) \approx$$

$$l =$$

$$\text{JaroWinkler}(s_1, s_2) \approx$$

$$t = \frac{\text{unequal positions in matching chars}}{2}$$

m_1	M	e	o	s	i	r	a
			$\neq$	$\neq$	$\neq$	$\neq$	
m_2	M	e	r	i	s	o	a

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$$|s_1| = 8$$

$$|s_2| = 12$$

$$w = 5$$

$$m = 7$$

$$t = 2$$

$$\text{Jaro}(s_1, s_2) \approx$$

$$l =$$

$$\text{JaroWinkler}(s_1, s_2) \approx$$

$$t = \frac{\text{unequal positions in matching chars}}{2}$$

m_1	M	e	o	s	i	r	a
			$\neq$	$\neq$	$\neq$	$\neq$	
m_2	M	e	r	i	s	o	a

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$$|s_1| = 8$$
$$|s_2| = 12$$
$$w = 5$$
$$m = 7$$
$$t = 2$$
$$\text{Jaro}(s_1, s_2) \approx$$
$$l =$$
$$\text{JaroWinkler}(s_1, s_2) \approx$$

$$\text{Jaro}(s_1, s_2) = \begin{cases} \frac{1}{3} \times \left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m} \right) & : m > 0 \\ 0 & : \text{otherwise} \end{cases}$$

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$$|s_1| = 8$$
$$|s_2| = 12$$
$$w = 5$$
$$m = 7$$
$$t = 2$$
$$\text{Jaro}(s_1, s_2) \approx 0.72$$
$$l =$$
$$\text{JaroWinkler}(s_1, s_2) \approx$$

$$\text{Jaro}(s_1, s_2) = \begin{cases} \frac{1}{3} \times \left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m} \right) & : m > 0 \\ 0 & : \text{otherwise} \end{cases}$$

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$$|s_1| = 8$$
$$|s_2| = 12$$
$$w = 5$$
$$m = 7$$
$$t = 2$$
$$\text{Jaro}(s_1, s_2) \approx 0.72$$
$$l =$$
$$\text{JaroWinkler}(s_1, s_2) \approx$$

$l =$ length of common prefix up to 4

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$$|s_1| = 8$$

$$|s_2| = 12$$

$$w = 5$$

$$m = 7$$

$$t = 2$$

$$\text{Jaro}(s_1, s_2) \approx 0.72$$

$$l = 2$$

$$\text{JaroWinkler}(s_1, s_2) \approx$$

l = length of common prefix up to 4

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$$|s_1| = 8$$
$$|s_2| = 12$$
$$w = 5$$
$$m = 7$$
$$t = 2$$
$$\text{Jaro}(s_1, s_2) \approx 0.72$$
$$l = 2$$
$$\text{JaroWinkler}(s_1, s_2) \approx$$

$$\text{JaroWinkler}(s_1, s_2) = \begin{cases} \text{Jaro}(s_1, s_2) + l \times 0.1 \times (1 - \text{Jaro}(s_1, s_2)) & : \text{Jaro}(s_1, s_2) \geq 0.7 \\ \text{Jaro}(s_1, s_2) & : \text{otherwise} \end{cases}$$

Jaro-Winkler Similarity

Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2	M	e	r	i	s	m	o	p	e	d	i	a

$$|s_1| = 8$$
$$|s_2| = 12$$
$$w = 5$$
$$m = 7$$
$$t = 2$$
$$\text{Jaro}(s_1, s_2) \approx 0.72$$
$$l = 2$$
$$\text{JaroWinkler}(s_1, s_2) \approx 0.78$$

$$\text{JaroWinkler}(s_1, s_2) = \begin{cases} \text{Jaro}(s_1, s_2) + l \times 0.1 \times (1 - \text{Jaro}(s_1, s_2)) & : \text{Jaro}(s_1, s_2) \geq 0.7 \\ \text{Jaro}(s_1, s_2) & : \text{otherwise} \end{cases}$$

Bounded Jaro-Winkler Similarity Based Search

- Given: query term s_1, terminology S_2, threshold θ
- Task: Find all terms $s_2 \in S_2$ with Jaro-Winkler similarity $\geq \theta$
- naive approach: similarity computation for each pair
  - $\rightarrow$ expensive task
  - $\rightarrow$ e.g. ≈ 0.7 s for single query on terminology with 1,000,000 terms[1]
  - $\rightarrow$ not suitable for interactive use cases with several queries[2]

[1] not parallelized on machine with two Intel Xeon Scalable 6140 18 Core 2,3 Ghz processors and 192 GB memory
[2] Nielsen 1993. Usability Engineering

Bounded Jaro-Winkler Similarity Based Search

Approach by Dreßler et al.[1]

- filter term pairs by lengths and character frequencies
  - → early stop similarity computation
  - → powerful approach for matching of two terminologies
- for search: still processes each pair
  - → outperformed by the naive approach
  - → not suitable for the use case

→ avoid processing of each pair

[1] Dreßler et al. 2017. “On the efficient execution of bounded Jaro-Winkler distances”

Efficient Bounded Jaro-Winkler Similarity Based Search

Search Tree

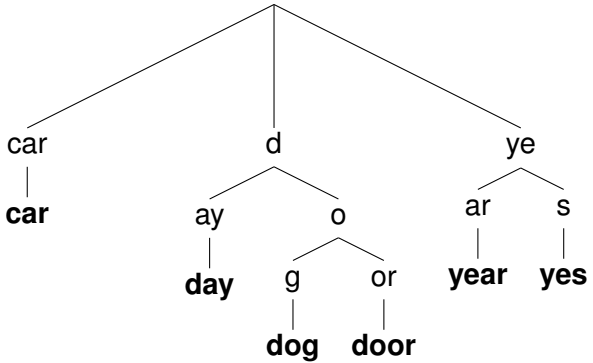
- avoid processing of each pair by storing terminology in *PATRICIA tree*

→ clustering of terms with common prefix

→ easy reuse of previous results

→ easy skipping of whole branches

- offline preparation of search tree
- break condition for search tree traversal required



Efficient Bounded Jaro-Winkler Similarity Based Search

Break Condition for Search Tree Traversal[1]

$$\text{JaroWinkler}(s_1, s_2) = \begin{cases} \text{Jaro}(s_1, s_2) + l \times p \times (1 - \text{Jaro}(s_1, s_2)) & : \text{Jaro}(s_1, s_2) \geq b_t \\ \text{Jaro}(s_1, s_2) & : \text{otherwise} \end{cases}$$

$$\Downarrow$$

$$\max_{s_2 \in S_2^*}(\text{JaroWinkler}) = \begin{cases} 1 - \left(1 - \max_{s_2 \in S_2^*}(\text{Jaro})\right) \times \left(1 - \max_{s_2 \in S_2^*}(l) \times p\right) & : \max_{s_2 \in S_2^*}(\text{Jaro}) \geq b_t \\ \max_{s_2 \in S_2^*}(\text{Jaro}) & : \text{otherwise} \end{cases}$$

[1] JaroWinkler, Jaro, l, m, and t depend on (s_1, s_2). S_2^* = Set of strings with prefix s_2^* and length $|s_2|$.

Efficient Bounded Jaro-Winkler Similarity Based Search

Break Condition for Search Tree Traversal[1]

$$\text{Jaro}(s_1, s_2) = \begin{cases} \frac{1}{3} \times \left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m} \right) & : m > 0 \\ 0 & : \text{otherwise} \end{cases}$$

$$\Downarrow$$

$$\max_{s_2 \in S_2^*}(\text{Jaro}) = \begin{cases} \frac{1}{3} \times \left(\frac{\max\limits_{s_2 \in S_2^*}(m)}{|s_1|} + \frac{\max\limits_{s_2 \in S_2^*}(m)}{|s_2|} + 1 - \frac{\min\limits_{s_2 \in S_2^*}(t)}{\max\limits_{s_2 \in S_2^*}(m)} \right) & : \max\limits_{s_2 \in S_2^*}(m) > 0 \\ 0 & : \text{otherwise} \end{cases}$$

[1] JaroWinkler, Jaro, l, m, and t depend on (s_1, s_2). S_2^* = Set of strings with prefix s_2^* and length $|s_2|$.

Efficient Bounded Jaro-Winkler Similarity Based Search

Break Condition for Search Tree Traversal[1]

$$\max_{s_2 \in S_2^*}(\text{JaroWinkler}) < \theta$$

$$\max_{s_2 \in S_2^*}(\text{JaroWinkler}) = \begin{cases} 1 - \left(1 - \max_{s_2 \in S_2^*}(\text{Jaro})\right) \times \left(1 - \max_{s_2 \in S_2^*}(l) \times p\right) & : \max_{s_2 \in S_2^*}(\text{Jaro}) \geq b_t \\ \max_{s_2 \in S_2^*}(\text{Jaro}) & : \text{otherwise} \end{cases}$$

$$\max_{s_2 \in S_2^*}(\text{Jaro}) = \begin{cases} \frac{1}{3} \times \left(\frac{\max_{s_2 \in S_2^*}(m)}{|s_1|} + \frac{\max_{s_2 \in S_2^*}(m)}{|s_2|} + 1 - \frac{\min_{s_2 \in S_2^*}(t)}{\max_{s_2 \in S_2^*}(m)} \right) & : \max_{s_2 \in S_2^*}(m) > 0 \\ 0 & : \text{otherwise} \end{cases}$$

[1] JaroWinkler, Jaro, l, m, and t depend on (s_1, s_2). S_2^* = Set of strings with prefix s_2^* and length $|s_2|$.

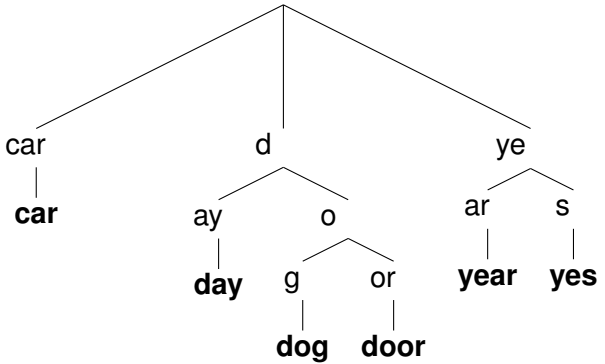
Efficient Bounded Jaro-Winkler Similarity Based Search

Non-Monotonicity of $\max_{s_2 \in S_2^*}(\text{JaroWinkler})$ **regarding** $|s_2|$

- changes of $|s_2|$ can increase or decrease $\max_{s_2 \in S_2^*}(\text{JaroWinkler})$
- only reuse results for equal $|s_2|$

$\rightarrow$ traverse search tree once for each $|s_2|$ value

$\rightarrow$ term lengths additionally stored at search tree nodes



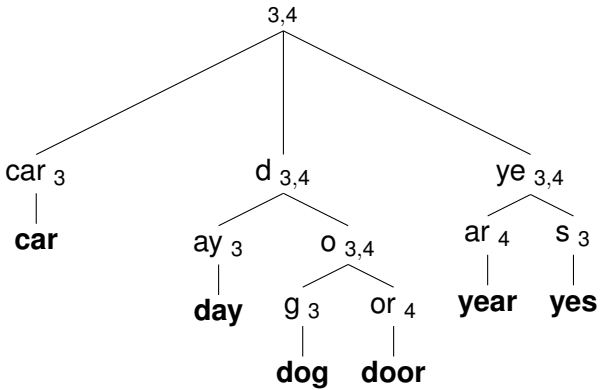
Efficient Bounded Jaro-Winkler Similarity Based Search

Non-Monotonicity of $\max_{s_2 \in S_2^*}(\text{JaroWinkler})$ **regarding** $|s_2|$

- changes of $|s_2|$ can increase or decrease $\max_{s_2 \in S_2^*}(\text{JaroWinkler})$
- only reuse results for equal $|s_2|$

$\rightarrow$ traverse search tree once for each $|s_2|$ value

$\rightarrow$ term lengths additionally stored at search tree nodes



Efficient Bounded Jaro-Winkler Similarity Based Search

Sequential Calculation Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2^*												

m_1	
m_2	

$$|s_1| = 8$$
$$|s_2| = 12$$
$$w = 5$$
$$m \leq 8$$
$$t \geq 0$$
$$l \leq 4$$
$$\text{JaroWinkler}(s_1, s_2) \leq 0.93$$

Efficient Bounded Jaro-Winkler Similarity Based Search

Sequential Calculation Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2^*	M											

m_1	
m_2	

$$|s_1| = 8$$
$$|s_2| = 12$$
$$w = 5$$
$$m \leq 8$$
$$t \geq 0$$
$$l \leq 4$$
$$\text{JaroWinkler}(s_1, s_2) \leq 0.93$$

Efficient Bounded Jaro-Winkler Similarity Based Search

Sequential Calculation Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2^*	M	e										

m_1	M
m_2	M

$$|s_1| = 8$$
$$|s_2| = 12$$
$$w = 5$$
$$m \leq 8$$
$$t \geq 0$$
$$l \leq 4$$
$$\text{JaroWinkler}(s_1, s_2) \leq 0.93$$

Efficient Bounded Jaro-Winkler Similarity Based Search

Sequential Calculation Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2^*	M	e	r									

m_1	M	e
m_2	M	e

$$|s_1| = 8$$
$$|s_2| = 12$$
$$w = 5$$
$$m \leq 8$$
$$t \geq 0$$
$$l \leq 4$$
$$\text{JaroWinkler}(s_1, s_2) \leq 0.93$$

Efficient Bounded Jaro-Winkler Similarity Based Search

Sequential Calculation Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2^*	M	e	r	i								

m_1	M	e	
m_2	M	e	r

$$|s_1| = 8$$
$$|s_2| = 12$$
$$w = 5$$
$$m \leq 8$$
$$t \geq 0$$
$$l \leq 2$$
$$\text{JaroWinkler}(s_1, s_2) \leq 0.91$$

Efficient Bounded Jaro-Winkler Similarity Based Search

Sequential Calculation Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2^*	M	e	r	i	s							

m_1	M	e		
m_2	M	e	r	i

$$|s_1| = 8$$
$$|s_2| = 12$$
$$w = 5$$
$$m \leq 8$$
$$t \geq 0$$
$$l \leq 2$$
$$\text{JaroWinkler}(s_1, s_2) \leq 0.91$$

Efficient Bounded Jaro-Winkler Similarity Based Search

Sequential Calculation Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2^*	M	e	r	i	s	m						

m_1	M	e			
m_2	M	e	r	i	s

$$|s_1| = 8$$
$$|s_2| = 12$$
$$w = 5$$
$$m \leq 8$$
$$t \geq 0$$
$$l \leq 2$$
$$\text{JaroWinkler}(s_1, s_2) \leq 0.91$$

Efficient Bounded Jaro-Winkler Similarity Based Search

Sequential Calculation Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2^*	M	e	r	i	s	m	o					

m_1	M	e				
m_2	M	e	r	i	s	o

$$|s_1| = 8$$
$$|s_2| = 12$$
$$w = 5$$
$$m \leq 8$$
$$t \geq 0$$
$$l \leq 2$$
$$\text{JaroWinkler}(s_1, s_2) \leq 0.91$$

Efficient Bounded Jaro-Winkler Similarity Based Search

Sequential Calculation Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2^*	M	e	r	i	s	m	o	p				

m_1	M	e				
m_2	M	e	r	i	s	o

$$|s_1| = 8$$
$$|s_2| = 12$$
$$w = 5$$
$$m \leq 8$$
$$t \geq 0$$
$$l \leq 2$$
$$\text{JaroWinkler}(s_1, s_2) \leq 0.91$$

Efficient Bounded Jaro-Winkler Similarity Based Search

Sequential Calculation Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2^*	M	e	r	i	s	m	o	p	e			

m_1	M	e	o	s	i	r
			$\neq$	$\neq$	$\neq$	$\neq$
m_2	M	e	r	i	s	o

$$|s_1| = 8$$
$$|s_2| = 12$$
$$w = 5$$
$$m \leq 7$$
$$t \geq 2$$
$$l \leq 2$$
$$\text{JaroWinkler}(s_1, s_2) \leq 0.78$$

Efficient Bounded Jaro-Winkler Similarity Based Search

Sequential Calculation Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2^*	M	e	r	i	s	m	o	p	e	d		

m_1	M	e	o	s	i	r
			$\neq$	$\neq$	$\neq$	$\neq$
m_2	M	e	r	i	s	o

$$|s_1| = 8$$
$$|s_2| = 12$$
$$w = 5$$
$$m \leq 7$$
$$t \geq 2$$
$$l \leq 2$$
$$\text{JaroWinkler}(s_1, s_2) \leq 0.78$$

Efficient Bounded Jaro-Winkler Similarity Based Search

Sequential Calculation Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2^*	M	e	r	i	s	m	o	p	e	d	i	

m_1	M	e	o	s	i	r
			$\neq$	$\neq$	$\neq$	$\neq$
m_2	M	e	r	i	s	o

$$|s_1| = 8$$
$$|s_2| = 12$$
$$w = 5$$
$$m \leq 7$$
$$t \geq 2$$
$$l \leq 2$$
$$\text{JaroWinkler}(s_1, s_2) \leq 0.78$$

Efficient Bounded Jaro-Winkler Similarity Based Search

Sequential Calculation Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2^*	M	e	r	i	s	m	o	p	e	d	i	a

m_1	M	e	o	s	i	r
			$\neq$	$\neq$	$\neq$	$\neq$
m_2	M	e	r	i	s	o

$$|s_1| = 8$$
$$|s_2| = 12$$
$$w = 5$$
$$m \leq 7$$
$$t \geq 2$$
$$l \leq 2$$
$$\text{JaroWinkler}(s_1, s_2) \leq 0.78$$

Efficient Bounded Jaro-Winkler Similarity Based Search

Sequential Calculation Example

	1	2	3	4	5	6	7	8	9	10	11	12
s_1	M	e	l	o	s	i	r	a				
s_2^*	M	e	r	i	s	m	o	p	e	d	i	a

m_1	M	e	o	s	i	r	a
			$\neq$	$\neq$	$\neq$	$\neq$	
m_2	M	e	r	i	s	o	a

$|s_1| = 8$

$|s_2| = 12$

$w = 5$

$m = 7$

$t = 2$

$l = 2$

$\text{JaroWinkler}(s_1, s_2) \approx 0.78$

Evaluation

How good is the approach?

Evaluation

- comparison of naive approach, approach by Dreßler et al. and our approach
- measurement parameter:
  - number of queries: 1, 10, 10^2, 10^3, 10^4, 10^5
  - number of terms: 1, 10, 10^2, 10^3, 10^4, 10^5, 10^6
  - threshold: 0.91, 0.95, 0.99
  - overlap: full, half, none
  - preparation: unprepared, prepared
- 20 measurements on 3 machines[1] = 60 measurements per configuration

[1] not parallelized, each two Intel Xeon Scalable 6140 18 Core 2,3 Ghz processors and 192 GB memory

Evaluation

Results

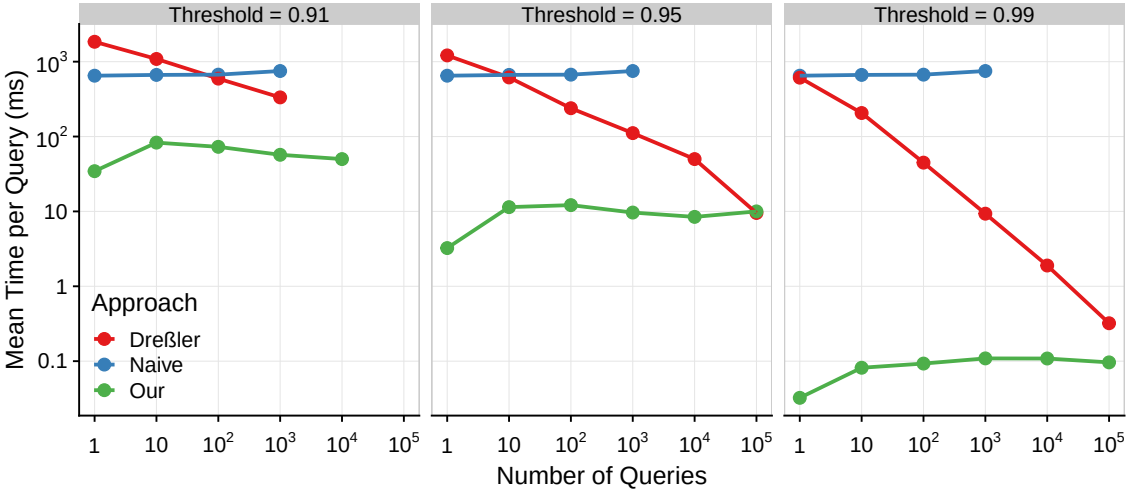
For 100 to 10^6 terms, threshold ≥ 0.91, and mixed full, half or no overlap:

- $\rightarrow$ unprepared terminology:
  statistically significant improved search efficiency for 10 to 10^3 queries
- $\rightarrow$ prepared terminology:
  statistically significant improved search efficiency for 1 to 10^3 queries

Evaluation

Results

Mean of measurements with prepared 10^6 terms and mixed full, half or no overlap.



Conclusion

Statistically significant improved efficiency of

Bounded Jaro-Winkler Similarity Based Search

Acknowledgments. Part of this work was funded by DFG in the scope of the LakeBase project within the Scientific Library Services and Information Systems (LIS) program. The computational experiments were performed on resources of Friedrich Schiller University Jena supported in part by DFG grants INST 275/334-1 FUGG and INST 275/363-1 FUGG. Many thanks to Frank Löffler for very helpful advice on the evaluation setup. Likewise many thanks to the three anonymous reviewers and the shepherd Ingo Schmitt for very helpful comments on earlier drafts of the manuscript.

Questions?

Implementation

github.com/fusion-jena/JaroWinklerSimilarity (Java, Apache 2.0)

Contact

jan-martin.keil@uni-jena.de
0000-0002-7733-0193
fusion.cs.uni-jena.de
@janmartinkeil

18. BTW 2019 Uni Rostock 600 Jahre

FUSION

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA