Semantically Weighted Similarity Analysis for XML-based Content Components

Jan Oevermann
jan.oevermann@dfki.de

Christoph Lüth
christoph.lueth@dfki.de

German Research Center for Artificial Intelligence

DocEng 2018, Halifax, 29.08.2018

Universität Bremen

Technical Documentation

- XML-based content components
  - Self-contained building blocks e.g. chapter-sized
  - Reuse, translation, aggregation, delivery
- Semantic XML information models
- Large databases of content components
- Product variants -> content variants

```
<descriptive nodeid="PI-70006536">
  <heading>Fuel Gas Requirements</heading>
  <descriptive_body>
    <paragraph>This Section defines […]
      <table>
      <row>
        <entry>
          <paragraph>Permissible range</parag
        </entry>
        <entry>
          <paragraph>
            <inlinedata>
              <si-value>
                <number>5</number>
                <unit>°C</unit>
              </si-value>
            </inlinedata>to
            <inlinedata>
              <si-value>
                <number>120</number>
                <unit>°C</unit>
              </si-value>
            </inlinedata>
          </paragraph>
```

Motivation

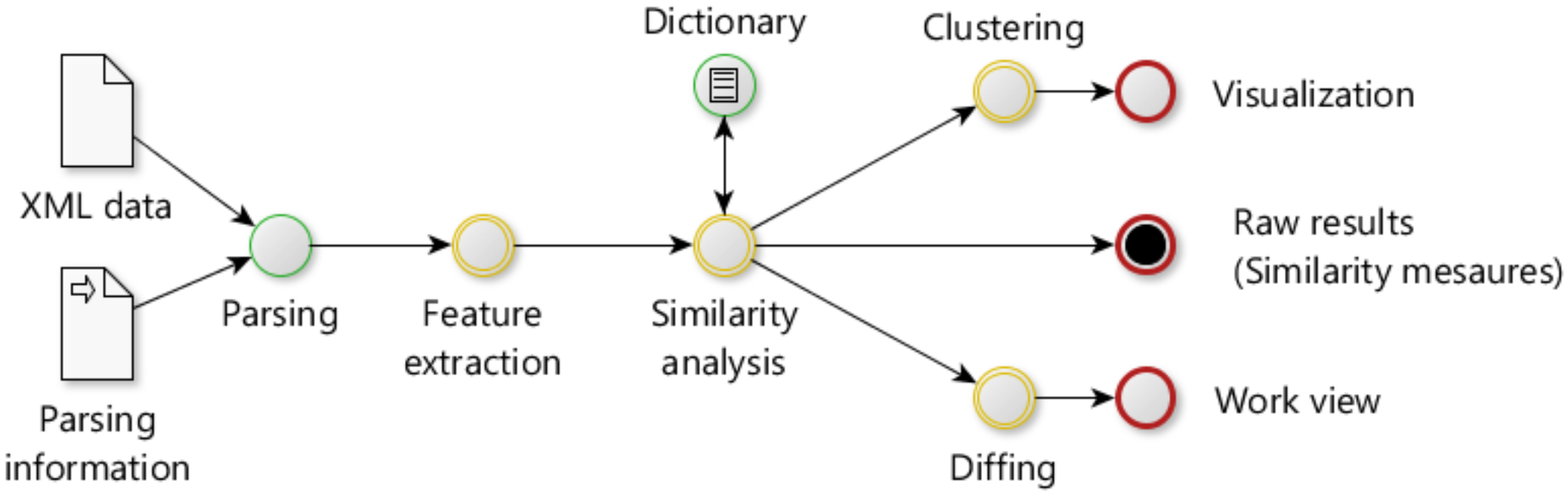
- Similar or duplicate content components
  - Document-based migration
  - Uncontrolled reuse / copying
  - Not checking / finding existing content

- Why is this bad?
  - Information retrieval / content delivery
    - high recall, low precision
  - Higher translation cost for variants
  - Time spent (re)writing existing content

Requirements & Implications

- Large amounts of content components
  - Computational efficient algorithm
- Simple similarity measure
  - Reliable against semantically similar differences
- (Non-)Detection of intentional variants
  - Weighting of semantically relevant text properties
- Quality assurance
  - UI for checking flagged relations

Architecture



Similarity analysis

- Similarity relations are symmetrical
- Total number of all relations (C) can grow rapidly

$$|C| = \frac{n * (n-1)}{2}$$

- Cosine similarity (s) for comparing vectors with extracted features
- Threshold for similarity measure to reduce total number of relations to check (r)

$$s = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| * |\vec{b}|}$$

Universität Bremen

Semantic similarity

expected similarity

A

```
<paragraph nodeid="a">This device is designed to work with
a voltage of <inlinedata><si-value><number>110</number>
<unit>V</unit></si-value></inlinedata> only.</paragraph>
```

B

```
<paragraph nodeid="b">This device is designed to work with
a voltage of <inlinedata><si-value><number>220</number>
<unit>V</unit></si-value></inlinedata> only.</paragraph>
```

C

```
<paragraph nodeid="c">This device works with a voltage of
<inlinedata><si-value><number>110</number><unit>V</unit>
</si-value></inlinedata> only.</paragraph>
```

Semantic weighting

- Extracted text from weighted elements treated separately
- Weighting artificially increases feature count by quantifier (q)
- Influences similarity in predictable ways
- Does not add to the complexity of the similarity analysis

	A	B	C
A		0.45	0.98
B	0.90		0.36
C	0.75	0.63	

semantic similarity

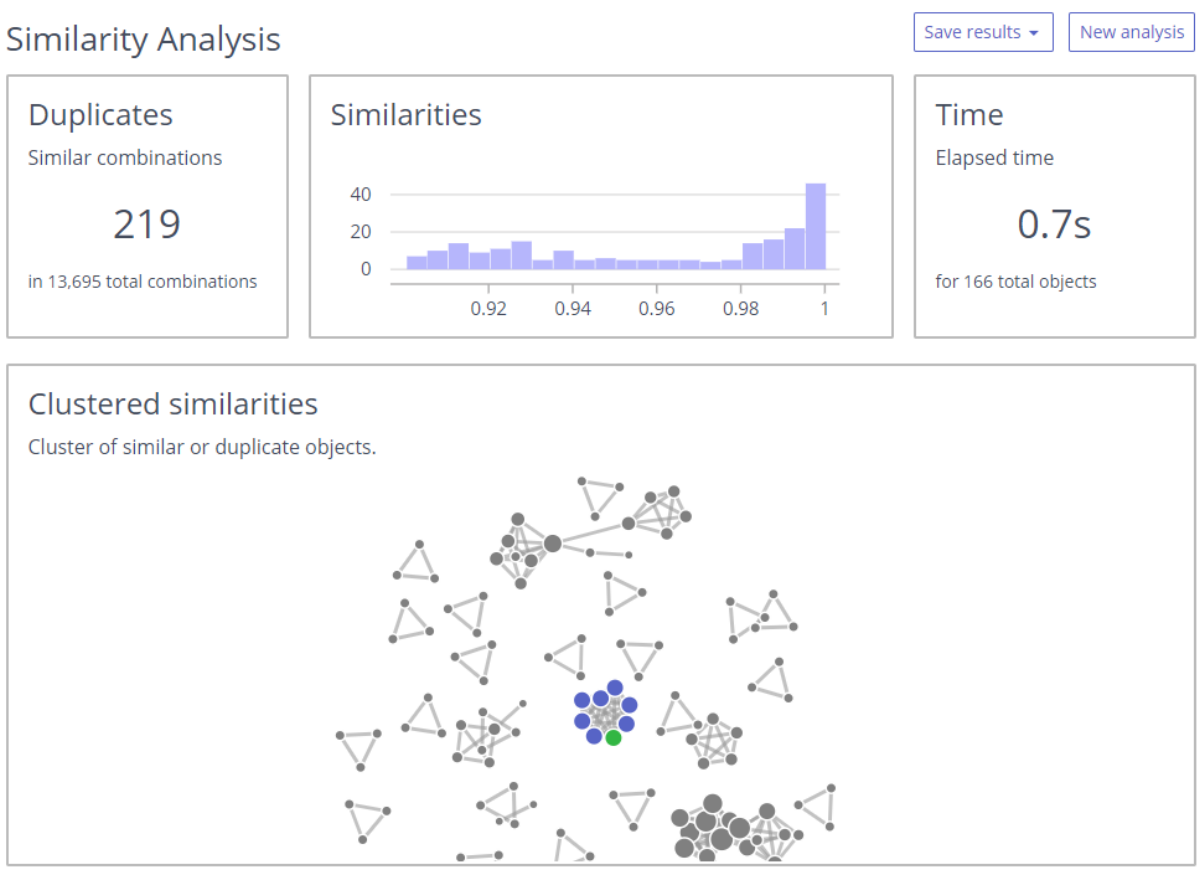
standard similarity

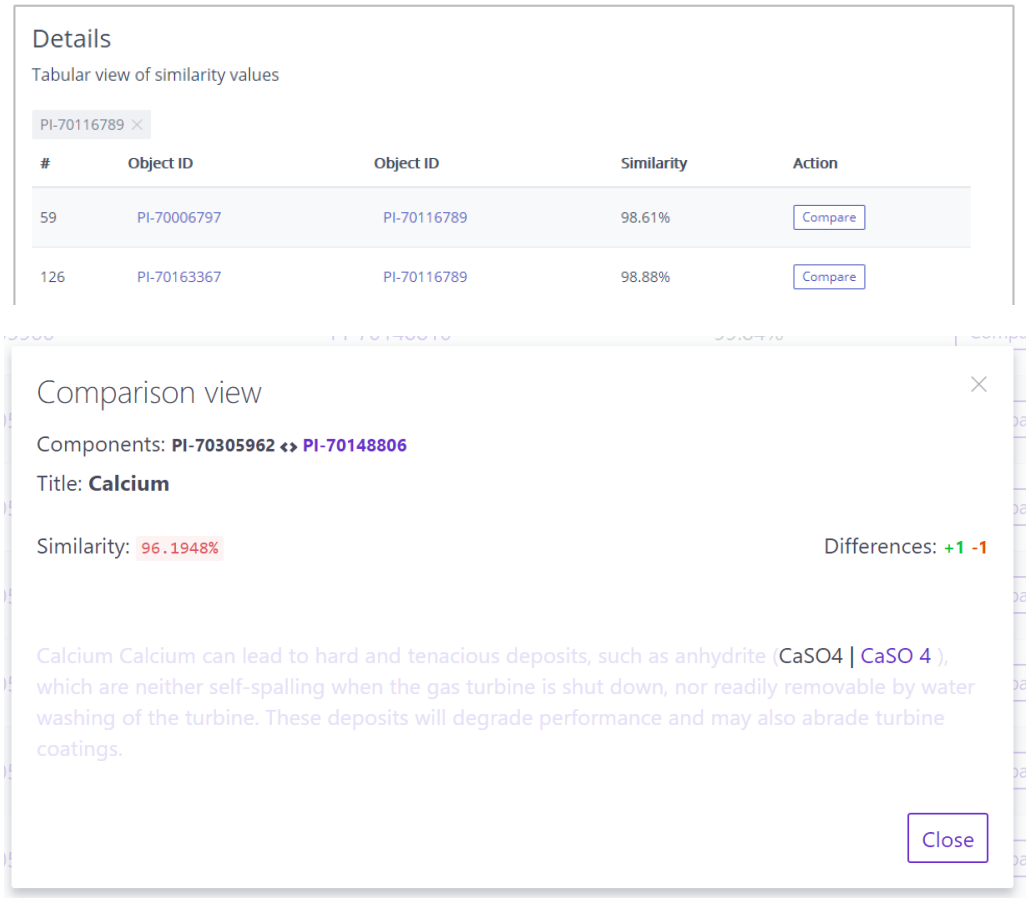
Implementation

- Implemented in JavaScript
- All processing is done client-side (browser), heavy calculations in own threads (web worker)
- Tested efficiency on standard hardware

Set	**units** (n)	**comb.** ($\|C\|$)	$\frac{\text{words}}{\text{unit}}$	**total t** [s]	$\frac{\text{t}}{\|C\|}$ [ms]
A	166	13,695	455.8	0.7	0.052
B	1,600	1,279,200	178.0	243.7	0.191
C	2,501	3,126,250	353.4	650.7	0.208
D	4,101	8,407,050	278.9	2,878.0	0.342

Workbench-like user interface





Universität Bremen

Similarity Analysis

Save results ▾ New analysis

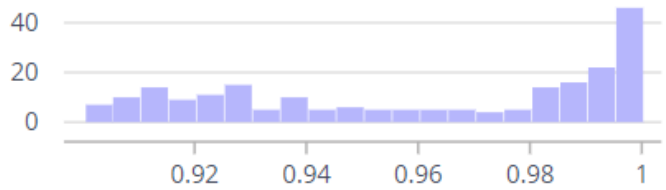
Duplicates

Similar combinations

219

in 13,695 total combinations

Similarities



Time

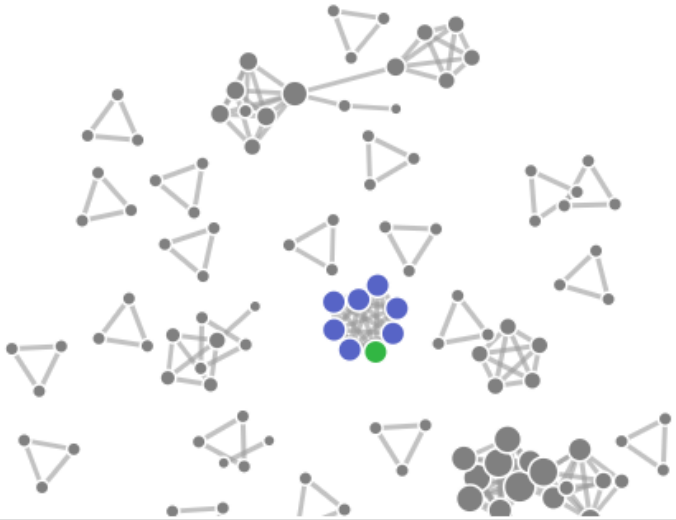
Elapsed time

0.7s

for 166 total objects

Clustered similarities

Cluster of similar or duplicate objects.



Universität Bremen

Details

Tabular view of similarity values

PI-70116789

#	Object ID	Object ID	Similarity	Action
59	PI-70006797	PI-70116789	98.61%	Compare
126	PI-70163367	PI-70116789	98.88%	Compare

Comparison view

Components: **PI-70305962 ↔ PI-70148806**

Title: **Calcium**

Similarity: 96.1948%

Differences: +1 -1

Calcium Calcium can lead to hard and tenacious deposits, such as anhydrite (CaSO4 | CaSO 4), which are neither self-spalling when the gas turbine is shut down, nor readily removable by water washing of the turbine. These deposits will degrade performance and may also abrade turbine coatings.

Close

Outlook & Conclusion

- RegEx or NER to in preprocessing to add XML tags
- Alternative similarity measures
- Integration with CCMS, give recommendations
- Research dependency to information model semanticity

- Simple method which can improve similarity results
- Real-world relevance through customer project with Siemens Energy (TecDoc Department)

Universität Bremen

Contact

Jan Oevermann

jan.oevermann@dfki.de

www.janoevermann.de

Code & Demo

github.com/j-oe/semsim

semsim.fastclass.de