Lesson4: Descriptive Modelling of Similarity of Text
Unit1: Similarity Measures

Rene Pickhardt

Introduction to Web Science Part 2
Emerging Web Properties

WeST
People and Knowledge Networks

Institute for Web Science and Technologies · University of Koblenz-Landau, Germany
Completing this unit you should …

• Know the properties of a similarity measure

• Be able to relate similarity and distance measures

• Know of two applications for modelling similarity
Similarity measures (definition & properties)

Given a Collection of text documents \( D \subseteq W^* \) for a finite set of words \( W = \{w_1, \ldots, w_N\} \)

\[ s : D \times D \longrightarrow \mathbb{R}^+ \] is called a similarity measure iff

- **Equal self-similarity** \( s(D_i, D_i) = s(D_j, D_j) \)
- **Symmetry** \( s(D_i, D_j) = s(D_j, D_i) \)
- **Maximality** \( s(D_i, D_i) \geq s(D_i, D_j) \)
Normalized similarity measures

Given a similarity measure \( s : D \times D \rightarrow \mathbb{R}^+ \)

We can deduce \( \tilde{s} : D \times D \rightarrow [0, 1] \) by setting

\[
\tilde{s}(D_i, D_j) = \frac{s(D_i, D_j)}{s(D_i, D_i)}
\]

Quiz:
- Why is this well defined?
- Do all the properties hold?
Connection to distance measures

Given a normalized similarity measure

\[ \tilde{s} : D \times D \longrightarrow [0, 1] \]

We can deduce a distance function by setting

\[ d(D_i, D_j) = -\log(\tilde{s}(D_i, D_j)) \]

Or the other way around:

\[ \Leftrightarrow \tilde{s}(D_i, D_j) = e^{-d(D_i, D_j)} \]
1st application: Ranking and querying

Given a query \( q \in W^* \) (or \( q \in D \) ?)

We can always assume that \( s \) can be extended to \( W^* \)

One can look at \( s(q, D_i) \forall D_i \in D \)

In particular at \( r_1 = \arg\max_{D_i \in D} \{ s(q, D_i) \} \)
We can iterate the process and create a ranking of a query based retrieval system

\[
\begin{align*}
    r_1 &= \arg\max_{D_i \in D} \{ s(q, D_i) \} \\
    r_2 &= \arg\max_{D_i \in D \setminus \{r_1\}} \{ s(q, D_i) \} \\
    r_3 &= \arg\max_{D_i \in D \setminus \{r_1, r_2\}} \{ s(q, D_i) \}
\end{align*}
\]

And so on for as many result documents as we want to retrieve.
2nd application: Recommender Systems

• Given a Document $D_j$

• Compute $s(D_i, D_j) \forall D_i \in D$

• And like before $r_1 = \arg\max_{D_i \in D \setminus \{D_j\}} \{s(D_i, D_j)\}$

• And iterate again for more results
Discussion

• Often natural similarity measures or natural distance measures occur

• Minimality becomes Maximality and vice versa

• You should get used to the fact that we and other people mix the terms (similarity and distance).

• Once the concept is understood you will do the same

• The omitted triangle inequality has better semantics for distance measures but won’t translates to similarities
Thank you for your attention!

Contact:
Rene Pickhardt
Institute for Web Science and Technologies
Universität Koblenz-Landau
rpickhardt@uni-koblenz.de
Copyright:

- This Slide deck is licensed under creative commons 3.0. share alike attribution license. It was created by Rene Pickhardt. You can use share and modify this slide deck as long as you attribute the author and keep the same license.

- [https://commons.wikimedia.org/wiki/File:Synoptic_word-for-word.png](https://commons.wikimedia.org/wiki/File:Synoptic_word-for-word.png) By Alecmconroy (Own work) [GFDL (http://www.gnu.org/copyleft/fdl.html) or CC BY 3.0 (http://creativecommons.org/licenses/by/3.0)], via Wikimedia Commons