Towards Named-Entity-based Similarity Measures: Challenges and Opportunities

Tom De Nies¹, Christian Beecks², Wesley De Neve¹,³, Thomas Seidl², Erik Mannens¹ and Rik Van de Walle¹

¹ Ghent University – iMinds – MMLab, Belgium
² RWTH Aachen University – DME Group, Germany
³ KAIST – Image and Video Systems Lab, Republic of Korea

{tom.denies, wesley.deneve, erik.mannens, rik.vandewalle}@ugent.be
{beecks, seidl}@informatik.rwth-aachen.de

4 CHALLENGES to improve existing document similarity measures through semantic awareness

1. ANNOTATION
Many techniques: categorization, topic detection, NER, linking, …

… it all boils down to disambiguation

bass vs. bass

Errors in disambiguation will result in less precise similarity measurement

2. SIMILARITY MEASURES
… for documents

Adapted traditional measures*
Documents must share at least one Named Entity

Adaptive distance-based measures**
No shared Named Entities needed

… to get meaningful values
[Examples: Jaccard, CF-IDF, TF-IS]

… if you know their distance
[Examples: EMD, SQFD, SMD]

… for individual Named Entities

ontology-based [1]

link-based [2]

shared-links-based [3]

3. LINKED DATA QUALITY
The LOD cloud still has a high number of missing links …
[Simonic, Rupnik and Siklaba, "Missing Properties in Linked Data Datasets" – lddminer.net]

… while its popularity has lead to spam and link pollution
[Hasnain et. al. “Spamming in Linked Data” @ COLD 2012]

4. LINKED DATA ACCESS
The LOD cloud offers a panorama of knowledge …

… which we view through a peephole (i.e., a SPARQL endpoint)

A “bag of words” has the advantage of always being up

<95% AVAILABILITY

Unfortunately, we can’t say the same about SPARQL endpoints …

So, alternative methods for reliable & scalable querying are needed

