Normalized Semantic Web Distance

Tom De Nies, Christian Beecks,
Fréderic Godin, Wesley De Neve, Grzegorz Stepien,
Dörthe Arndt, Laurens De Vocht, Ruben Verborgh,
Thomas Seidl, Erik Mannens and Rik Van de Walle

tom.denies@ugent.be

@TomDeNies

UNIVERSITEIT GENT

Contents

Introduction to the problem

Normalized Web Distance

Our approach: Normalized Semantic Web Distance

Evaluation & Discussion

iMinds

RWTH AACHEN UNIVERSITY

The Problem

We want to **quantify the (dis)similarity** or ***distance*** between things in the world

For indexing, retrieval, clustering, …

& because we can!

Humans can make a relative comparison…

… but it's much harder to put a number on it!

Normalized Web Distance

Approximates the (non-computable)
Normalized Information Distance

Based on statistics gathered through a **Web search engine**

→ the search engine index **"compresses" the information** contained within a concept to its **number of occurrences** on the Web

The theory: *If two concepts occur together frequently, their distance is smaller than when they don't*

Normalized Web Distance (continued)

To implement this, we use these frequency functions:

f(x) = the # of search results for a **concept x**

f(y) = the # of search results for a **concept y**

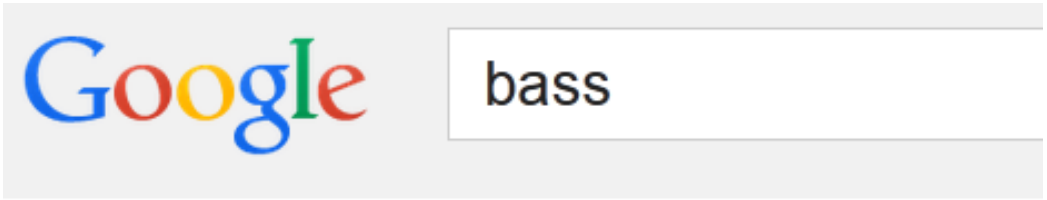
f(x,y) = the # of search results for **x and y together**

N = the total number of Web pages indexed by search engine

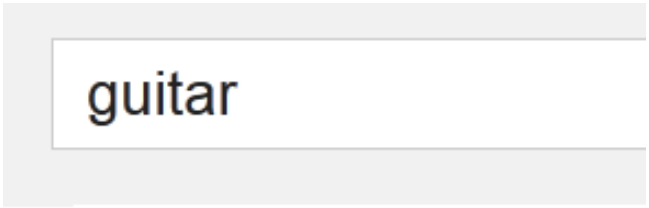
(can we even know this?)

$$NWD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}}$$

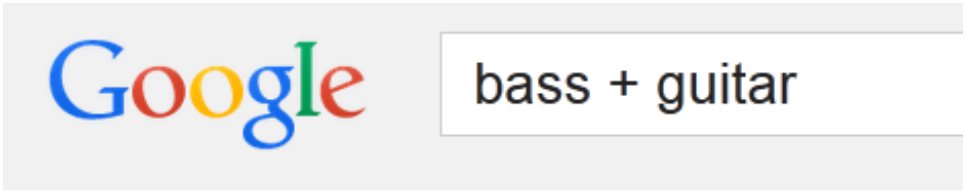
Example: Normalized Google Distance



About 29,500,000 results



About 30,900,000 results



About 2,810,000 results

$$NGD(bass, guitar) = \frac{\log 30900000 - \log 2810000}{\log(10 * 10^9) - \log 29500000}$$

$$= \frac{7.49 - 6.45}{10 - 7.47} \approx \mathbf{0.411}$$

This has some drawbacks

Availability – Access - Reproducibility

Google might not always be there… or it might change.
(search engine queries are not standardized)

You could use a different search engine (e.g., Bing).

Results would still vary at any moment in time.

Semantic Awareness

vs.

Our approach: Normalized Semantic Web Distance

Could we use the **principles of the NWD**, but with the **semantic awareness** of a **knowledge graph**?

Core principle: *If two concepts are used to describe the same things, their distance is smaller than when they're not.*

Normalized Semantic Web Distance

Knowledge graph (V, T):

nodes V, triples $T \subseteq V \times P \times V$, predicates P

$V_{in}(x)$ = **# nodes linking to x**

$V_{out}(x)$ = **# nodes linked to by x**

$V_{all}(x)$ = **# nodes that link to x OR that x links to**

Use the cardinality of these sets as frequency functions!

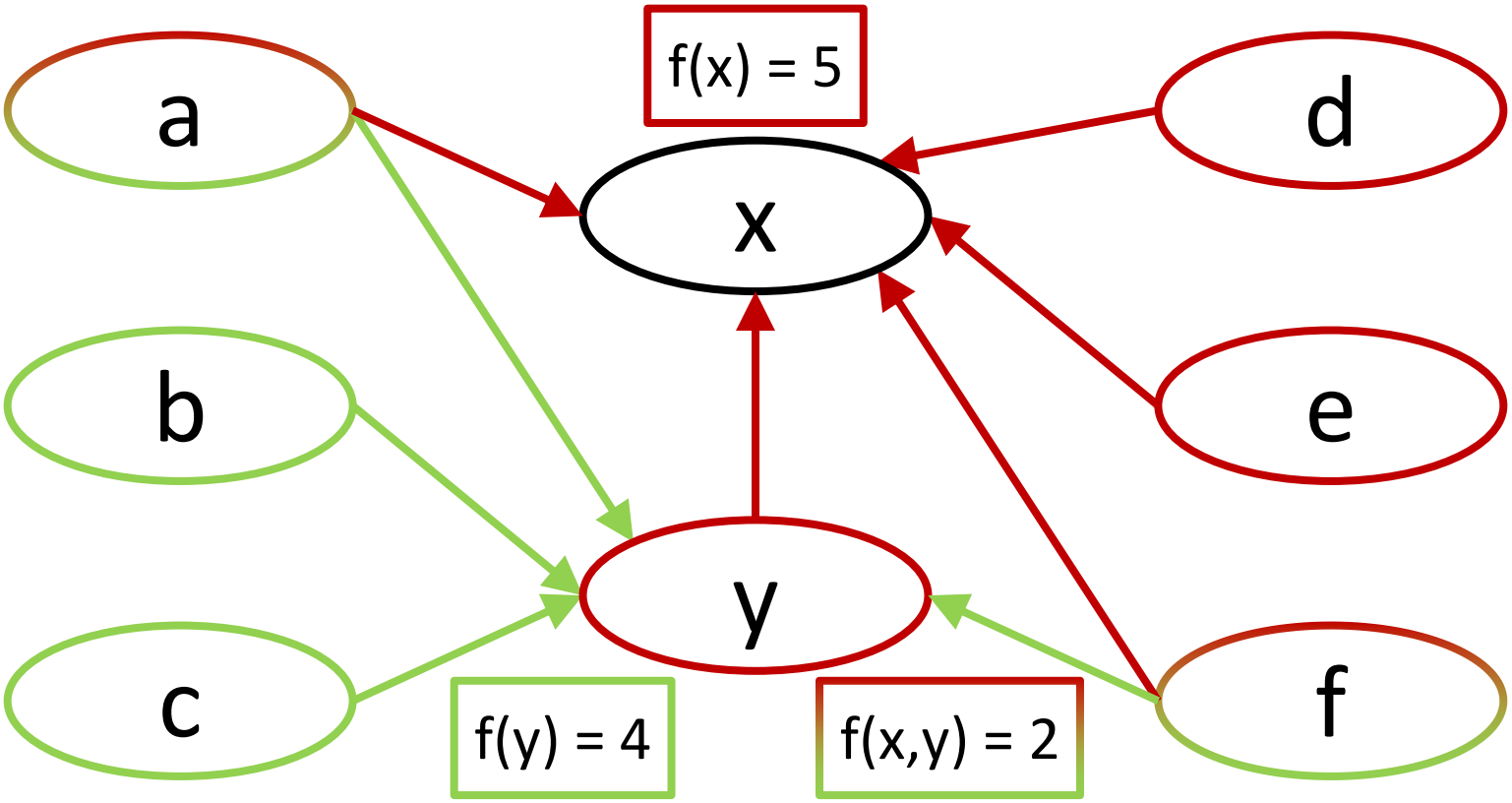
Normalized Semantic Web Distance

$$\lambda \in \{in, out, all\}:$$

$$NSWD_{\lambda}(x, y) = \frac{\max\{log|V_{\lambda}(x)|, log|V_{\lambda}(y)|\} - log|V_{\lambda}(x) \cap V_{\lambda}(y)|}{log|V| - min\{log|V_{\lambda}(x)|, log|V_{\lambda}(y)|\}}$$

Easily implementable using SPARQL COUNT DISTINCT queries!

Example for incoming links



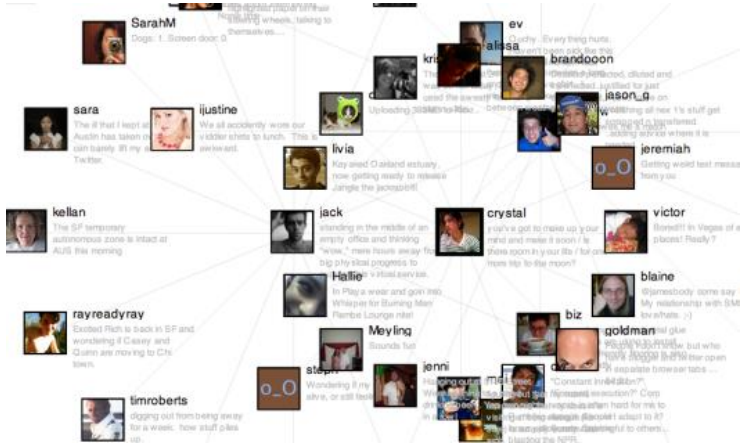
$$NSWD_{in}(x,y) = \frac{\log 5 - \log 2}{\log 1000 - \log 4} \approx \mathbf{0.16595}$$

Size of the knowledge graph

Applicable to any graph…

…intended for representing knowledge

…or for something else?



Evaluation

On **Miller-Charles MC30** set
30 term pairs + human judgment of similarity

car – automobile (3.92)

gem – jewel (3.84)

cemetery – woodland (0.95)

…

→ measure correlation

Term-pairs need to be **disambiguated**. 3 strategies tested:

- M: Manual (human judgment)
- C: Count-based (highest $|\boldsymbol{V}_{\lambda}|$)
- S: Similarity-based (smallest resulting distance between term-pair)

Converting Distance to Similarity

MC30 assessments reflect similarity, not dissimilarity

→ How do we convert NSWD to similarity measure?

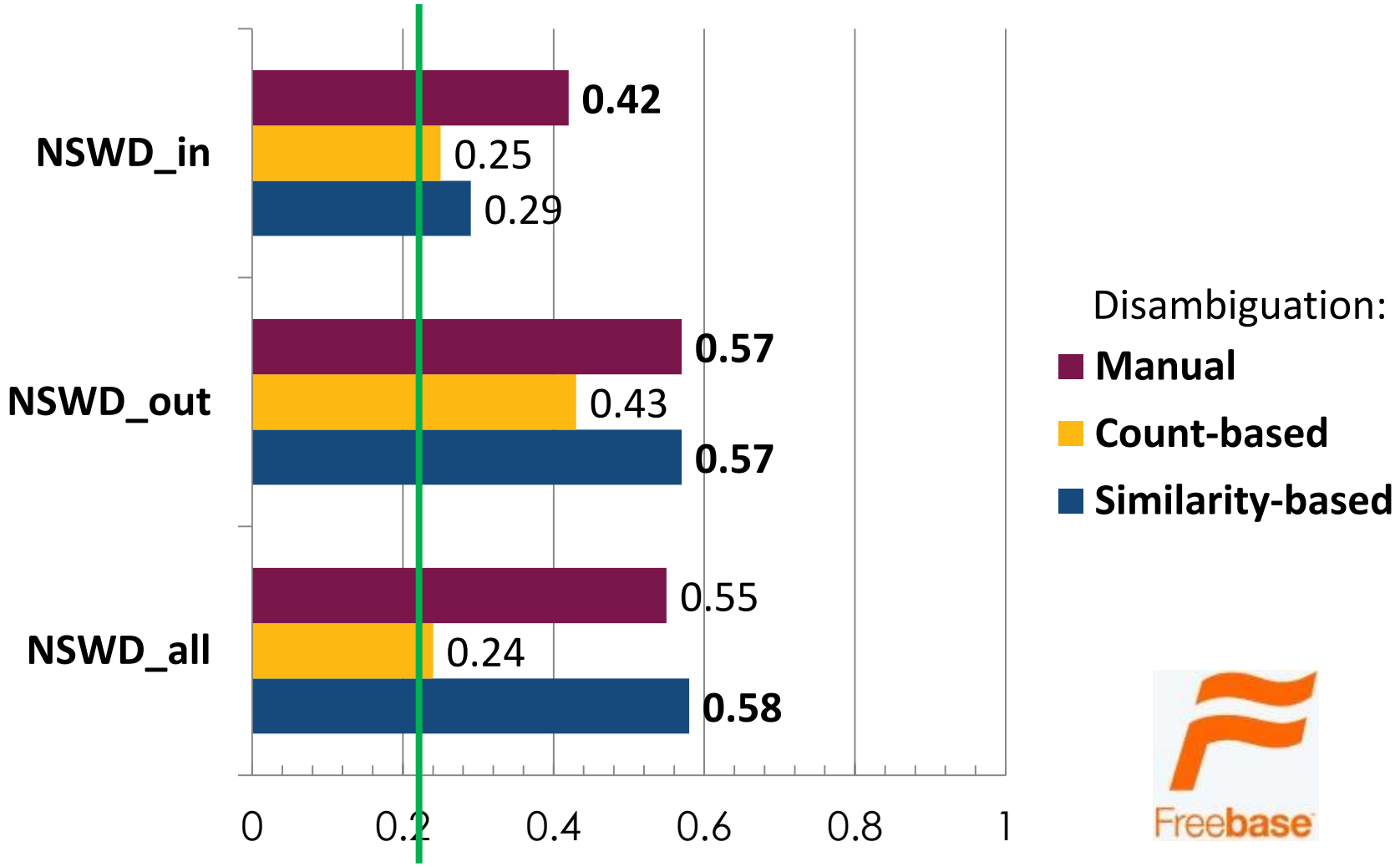
Using 1- $NSWD/NSWD_{max}$ (or a more complex scaling)

$$NSWD_{max} = \frac{log(\left\lfloor \frac{|V|}{2} \right\rfloor + 1)}{log\,|V| - log\left\lceil \frac{|V|}{2} \right\rceil}$$

For example, for a knowledge graph with **1000 nodes**:

$NSWD_{max}$ = log 501 / (log 1000 – log 500) ≈ 8.9686

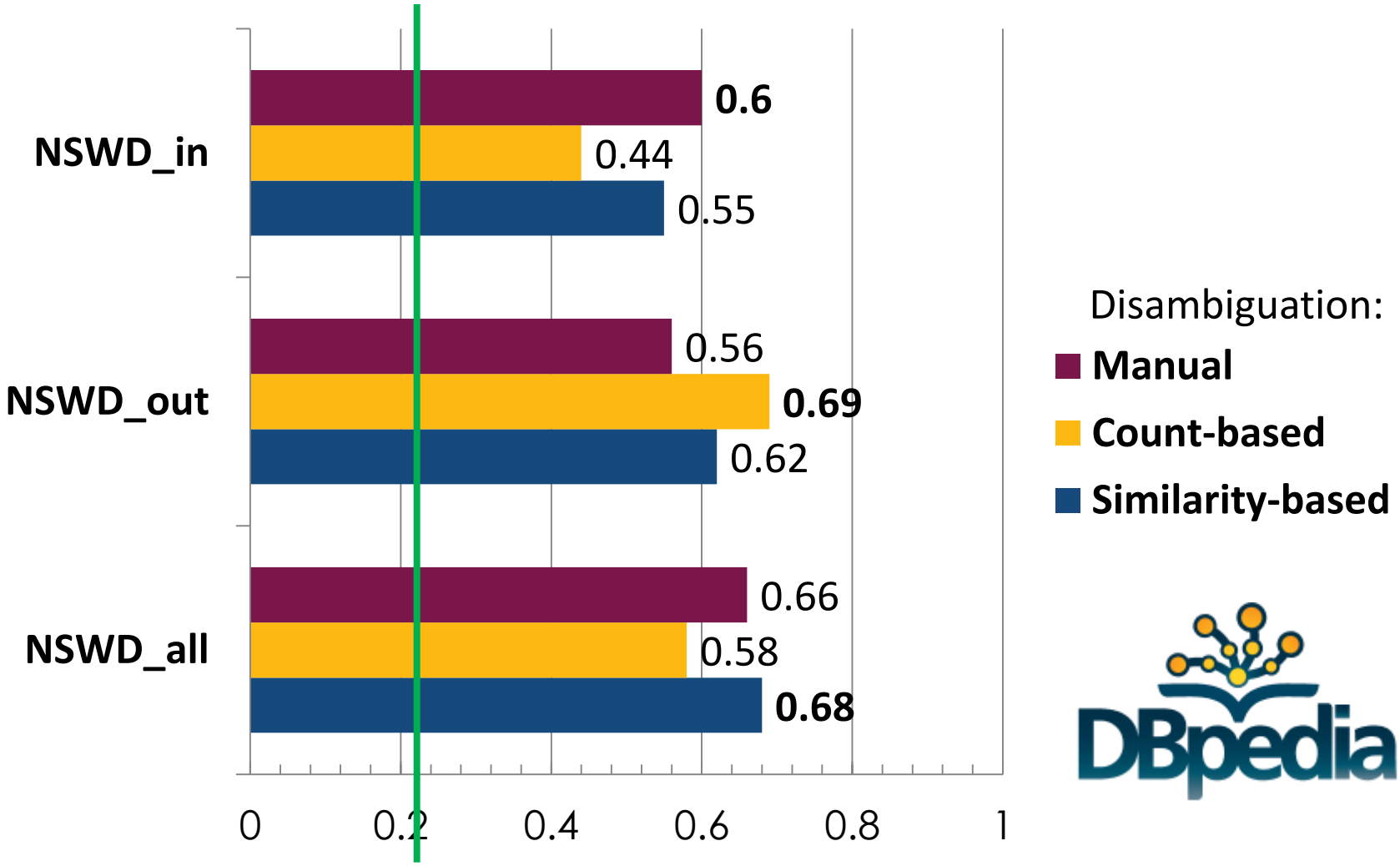
Evaluation: Correlation on Freebase Graph



NWD (Bing): 0.23

Freebase

Evaluation: Correlation on DBpedia Graph



Other approaches from literature

Wikipedia Link-based Measure (Milne & Witten, 2008)

→ combines NGD + traditional tf*idf weights

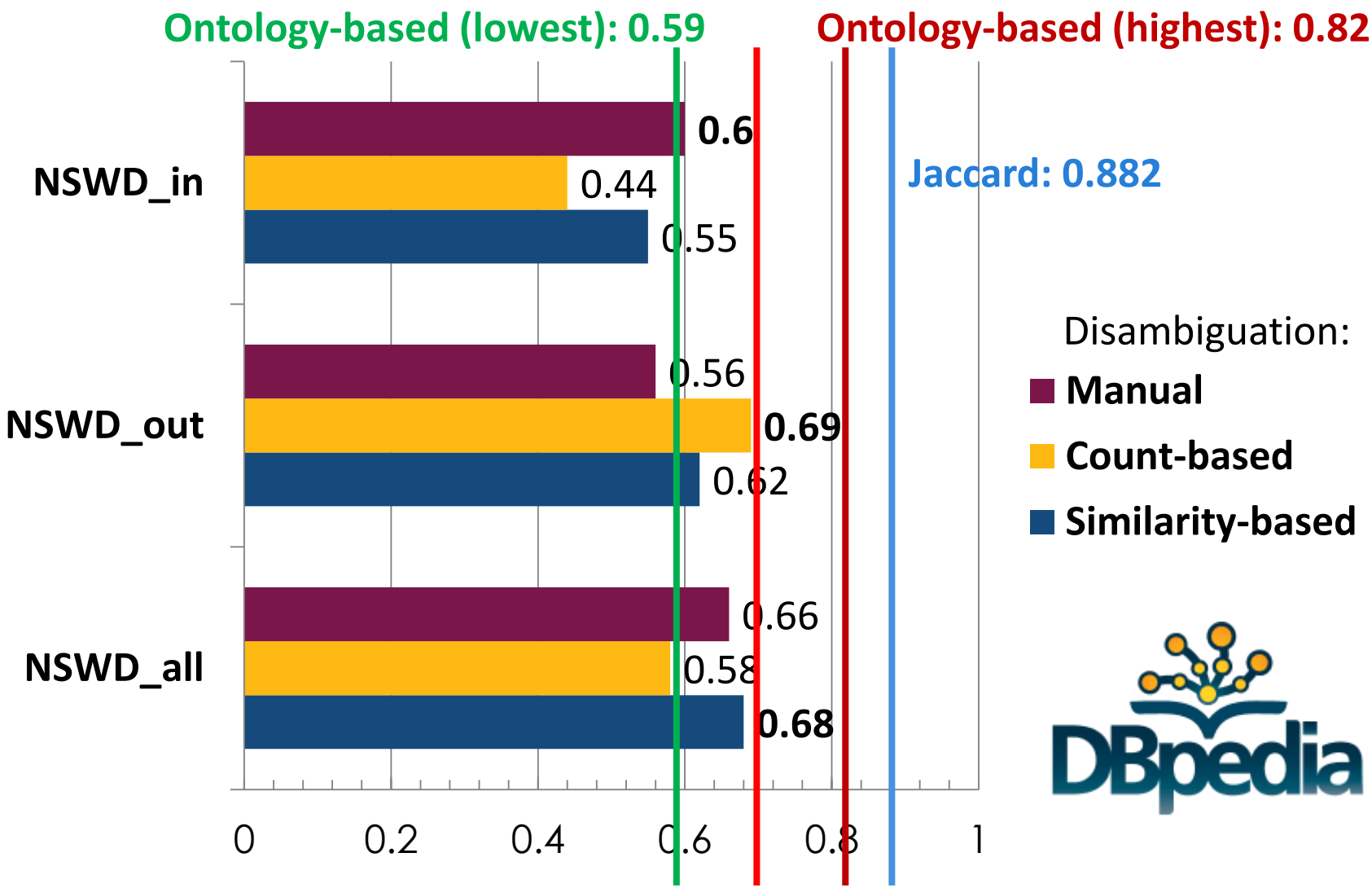
Ontology-based approaches (Hlioutakis et al., 2006 Sanchez et al., 2012)

→ distance within the concept hierarchy

Jaccard Distance (Kulkarni et al., 2009)

→ shared features vs. total features

Comparison to other approaches



Discussion

We **did not beat the existing approaches**.

However, we did achieve a **clear positive correlation**, and **outperformed the NWD baseline**

… on a **dataset not really suitable** for our approach!
(ambiguity was purposely built in the Miller-Charles dataset, but it was the only one available for external comparison)

Ambiguity in MC30 & Knowledge Graph Quality

Knowledge graphs don't always represent the knowledge we are trying to grasp

Examples:

dbpedia:journey & *dbpedia:voyage*

→ many disambiguation options, but none capture the intended meaning!

→ actually *dbpedia:travel*!

"lad", "madhouse" → no equivalent on Freebase

This leads to many variations in the results!

Can we evaluate something else?

Semantic Awareness

We designed our approach to have semantic awareness

vs.

Can we illustrate that claim?

We created a set of 20 concept pairs to illustrate this

Concept Pairs

concept1	concept2	concept3
:Pear	:Apple	:Apple_Inc
:Trout	:Bass_(fish)	:Bass_guitar
:Cat	:Jaguar	:Jaguar_Cars
:Cat	:Mouse	:Mouse_(computing)
:Automobile	:Bus	:Bus_(computing)
:Indonesia	:Java	:Java_(programming_language)
:Lion	:Tiger	:Tiger_(Danish_store)
:Musical_theatre	:Broadway_(play)	:Broadway_(Manhattan)
:Bird	:Crane_(bird)	:Crane_(machine)
:Bass_guitar	:String_(music)	:String_(physics)

We want:
NSWD(concept1, concept2) < NSWD(concept1, concept3)

Results for DBpedia

concept1	concept2	concept3
:Pear	:Apple	:Apple_Inc
:Trout	:Bass_(fish)	:Bass_guitar
:Cat	:Jaguar	:Jaguar_Cars
:Cat	:Mouse	:Mouse_(computing)
:Automobile	:Bus	:Bus_(computing)
:Indonesia	:Java	:Java_(programming_language)
:Lion	:Tiger	:Tiger_(Danish_store)
:Musical_theatre	:Broadway_(play)	:Broadway_(Manhattan)
:Bird	:Crane_(bird)	:Crane_(machine)
:Bass_guitar	:String_(music)	:String_(physics)

We want:
NSWD(concept1, concept2) < NSWD(concept1, concept3)

Results for Freebase

concept1	concept2	concept3
:Pear	:Apple	:Apple_Inc
:Trout	:Bass_(fish)	:Bass_guitar
:Cat	:Jaguar	:Jaguar_Cars
:Cat	:Mouse	:Mouse_(computing)
:Automobile	:Bus	:Bus_(computing)
:Indonesia		:Java_(programming_language)
:Lion	:Tiger	:Tiger_(Danish_store)
:Musical_theatre	:Broadway_(play)	:Broadway_(Manhattan)
:Bird	:Crane_(bird)	:Crane_(machine)
:Bass_guitar	:String_(music)	:String_(physics)

$NSWD_{in}$(:Automobile, :Bus) = 0.34

>

$NSWD_{in}$(:Automobile, :Bus_(computing)) = 0.32

We want:

NSWD(concept1, concept2) < NSWD(concept1, concept3)

Next steps

Use a different, **unambiguous evaluation corpus**
(e.g., Hulpus et al., 2015 ?)

What **aspects of the knowledge graph** play a part?

- Quality
- Connectivity
- Size
- Domain-specificity
- …

Can we measure this with NSWD?

Integrate into **adaptive dissimilarity measures for documents**

UNIVERSITEIT GENT

Conclusion

We introduced a **new way of measuring distance in a knowledge graph**, advancing the idea of Normalized Web Distance

We **evaluated** our approach using the **Freebase and DBpedia** graphs on an **established benchmark**, and compared it with the Normalized Web Distance using the **Bing** search engine

While our approach **did not outperform others** from literature, it **did outperform the NWD baseline**, and showed **clear positive correlations** with human judgment.

We **illustrated the semantic awareness** of our approach