Normalized Semantic Web Distance

Tom De Nies[1(✉)], Christian Beecks[2], Fréderic Godin[1], Wesley De Neve[1,3], Grzegorz Stepien[2], Dörthe Arndt[1], Laurens De Vocht[1], Ruben Verborgh[1], Thomas Seidl[2], Erik Mannens[1], and Rik Van de Walle[1]

[1] iMinds – Data Science Lab, Ghent University, Ghent, Belgium
{tom.denies,frederic.godin,wesley.deneve,dorthe.arndt,laurens.devocht,
ruben.verborgh,erik.mannens,rik.vandewalle}@ugent.be
[2] DME Group, RWTH Aachen University, Aachen, Germany
{beecks,grzegorz.stepien,seidl}@informatik.rwth-aachen.de
[3] IVY Lab, KAIST, Daejeon, Republic of Korea

Abstract. In this paper, we investigate the Normalized Semantic Web Distance (NSWD), a semantics-aware distance measure between two concepts in a knowledge graph. Our measure advances the Normalized Web Distance, a recently established distance between two textual terms, to be more semantically aware. In addition to the theoretic fundamentals of the NSWD, we investigate its properties and qualities with respect to computation and implementation. We investigate three variants of the NSWD that make use of all semantic properties of nodes in a knowledge graph. Our performance evaluation based on the Miller-Charles benchmark shows that the NSWD is able to correlate with human similarity assessments on both Freebase and DBpedia knowledge graphs with values up to 0.69. Moreover, we verified the semantic awareness of the NSWD on a set of 20 unambiguous concept-pairs. We conclude that the NSWD is a promising measure with (1) a reusable implementation across knowledge graphs, (2) sufficient correlation with human assessments, and (3) awareness of semantic differences between ambiguous concepts.

1 Introduction

The goal of semantic distance and/or similarity measures is to mimic the human assessment of distance between two concepts. However, humans usually do not explicitly quantify the distance between concepts – at least not consciously – which makes the development and evaluation of semantic distances a challenging task. These measures play an important role on the Web, particularly when it comes to indexing, semantic search, and information retrieval. Despite being an intensely researched field for the past decades, traditional, plain-text-based similarity measures are still the dominant norm in practical scenarios [9]. This is unfortunate because most of these have little to no semantic awareness since they work with syntactic information instead of machine-interpretable data.

We argue that increasing the semantic awareness of similarity measures is possible by making use of machine-interpretable concepts that are unambiguously linked to a resource on the Web [6]. Determining the distance between

© Springer International Publishing Switzerland 2016
H. Sack et al. (Eds.): ESWC 2016, LNCS 9678, pp. 69–84, 2016.
DOI: 10.1007/978-3-319-34129-3_5

individual concepts facilitates the application of more complex similarity measures such as Earth Mover's Distance or Signature Quadratic Form Distance [1] on Web documents that are modeled by the concepts they contain. To this end, we investigate the properties of the Normalized Semantic Web Distance (NSWD), an inter-concept distance we introduced [7] based on the statistics of the context of the entities in the knowledge graph where they are represented. In this paper, we provide additional theoretical fundamentals of the NSWD, as well as insights into its computation and implementation. Additionally, we further evaluate our approach on the DBpedia knowledge graph as well as Freebase, and verify its semantic awareness on a set of 20 unambiguous concept pairs.

2 Background Knowledge and Motivation

The distance measure we introduce in this paper is based on the Normalized Information Distance (NID), introduced in [15]. The NID is defined using the so called *Kolmogorov Complexity*. Formally, the Kolmogorov Complexity of a binary string x is defined as the length of the shortest program p with $U(p) = x$ for a fixed universal prefix Turing machine U [14]. Informally, it can be seen as the length of the maximally compressed version of x and in a sense reflects the amount of information encoded in x. Unfortunately, the Kolmogorov Complexity is non-computable, and thus needs to be approximated. This approximation is strongly affected by the representation of objects, either in *literal* or *non-literal* form. The literal form contains the object itself. A song or a novel, for example, can be provided as a binary file containing the actual song or novel. The non-literal form consists of a (possibly ambiguous) name referencing the actual object. Abstract concepts like "love" or "beauty" can only be provided by their non-literal name. This corresponds to the vision of information and non-information resources described in the original World Wide Web architecture [12].

If the input for the NID is given in **literal form**, we can approximate the Kolmogorov Complexity by the size of the output of a state of the art compression algorithm. This principle was introduced as the Normalized Compression Distance (NCD) [3]. However, in this paper we focus on **non-literal** objects, and we need a different approximation. In [4], it was shown that for a frequency function f, the NID can be approximated as in Definition 1.

Definition 1. *For two binary strings x and y, let $f(x,y) \in \mathbb{N}_0$ be a frequency function for which $N := \sum_{x,y} f(x,y) < \infty$ and $f(x) := f(x,x)$ holds. f resembles how common x and y occur together in a certain context. The NID can now be approximated as:*

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

The challenge is to define a reasonable frequency function f. A well known instance is the so called Normalized Web Distance (NWD) [4]. It employs the statistics returned by an arbitrary Web search engine to calculate f, where the

frequency $f(x)$ of a term x is the number of indexed pages mentioning x. The basic principle of the NWD is: if two terms occur together almost as often as they do separately, their semantic distance is likely to be low. More formally:

Definition 2. *Let W be the set of pages indexed by an arbitrary search engine able to return the (approximate) number of indexed pages containing a certain search term. For each search term x, let $\mathbf{X} \subseteq W$ denote the set of pages containing x. For two search terms x and y, the* Normalized Web Distance *corresponds to the NID_f from Definition 1* and the following frequency function f [4]:

$$f(x) := |\mathbf{X}|$$
$$f(x, y) := |\mathbf{X} \cap \mathbf{Y}|$$

Note that search engines usually only estimate $f(x)$ and do not explicitly exclude duplicate pages from the search result. Furthermore, due to the volume of stored and indexed Web pages, a search engine cannot compute N as exactly as in Definition 1. However, since it merely serves as a scaling factor in the NID_f equation, it can be set to an arbitrary value $\geq |W|$ (this constraint ensures that the NWD is always non-negative).

Any search engine can be used for the NWD and *different engines* usually return *different results*. This approach is semantically unaware of the meaning of the input terms, as well as of the (human-understandable) Web pages returned by the search engine. The word "Java", for example, can either refer to the Indonesian island, or to the programming language. This means that calculating the NWD from the word "Java" to "Indonesia" leads to the same distance for either meaning, regardless of which the user intended. This is exactly the problem we are tackling in this paper: comparing unambiguous concepts instead of ambiguous lexical terms. While other approaches exist to achieve this (as seen in Sect. 3), none are based on the NID, which makes them less related to the way we represent knowledge. Since the essential foundations of the Semantic Web include an accurate representation of knowledge, it also makes sense to create a semantic distance that has roots in the same foundations. In Sect. 3, we highlight a number of known implementations of the NWD, as well as various other semantic distance or similarity measures relevant to our approach.

3 Related Work

The first instance of the NWD was the Normalized Google Distance (NGD), which made use of the Google search engine, and achieved up to 87 % agreement with human assessments [5]. However, due to changes to the functionality of the AND operator in the Google search engine [24], the returned page counts are no longer trustworthy. In our evaluation, we use the Bing search engine as an alternative, resulting in the Normalized Bing Distance (NBD) [10]. In fact, the NGD paper clearly states that other data sources than the World Wide Web, such as a dictionary, could be used to calculate the NGD [5]. This further reinforces our intuition of adapting the NWD to a graph-structured database.

The NGD has been used as an inspiration for a link-based similarity measure before, namely the Wikipedia Link-based Measure (WLM) [18]. WLM adapts the traditional TF-IDF-vector-based approach as well as the NGD approach to exploit the link structure of Wikipedia. When only considering the incoming-links-part of WLM, the principle is comparable to that of the NSWD, with the exception of WLM being specific to Wikipedia, whereas the NSWD is more generally applicable.

In previous work, we conducted a preliminary experiment with an approach we named the *Normalized Freebase Distance (NFD)* [10]. The NFD served as a first specific implementation of the concept of the NSWD. The promising results inspired us to generalize the concept to all knowledge graphs on the Semantic Web in [7]. We expand significantly upon the evaluation provided in these previous papers, as well as on the theoretical foundations for the NSWD.

The Jaccard similarity is one of the most efficient measures for semantic relatedness [2,13]. Unlike the Jaccard similarity, the Jaccard distance (inverse similarity) is a valid heuristic for e.g., a pathfinding algorithm. For example, the "Everything is Connected Engine" (EiCE) [8] uses a distance metric based on the Jaccard similarity for pathfinding. It applies the measure to estimate the similarity between two nodes and to assign a random-walk based weight, which ranks less popular resources higher, thereby guaranteeing that paths between resources prefer specific relations over general ones [19].

The Linked Data Semantic Distance (LDSD) [22] also partially relies on shared links. Similar to our proposed approach, the LDSD is extensible for specific domains, as shown in [25], where it is extended to model the similarity of human behavior processes.

An important category of semantic relatedness measures contains those that measure the distance between two concepts in the context of their concept hierarchy or ontology. Hliaoutakis et al. [11] provide a comparison of 11 such semantic relatedness measures, tested using WordNet[1] and MeSH[2]. More recently, Sanchez et al. [23] provide an overview over ontology-based semantic similarity measures, including a newly proposed one of their own. The correlation of these measures with popular benchmarks for word relatedness ranges between 0.7 and 0.86. For more details, we refer to the aforementioned surveys.

The Semantic Connectivity Score [21] is a measure to quantify the connectivity between two concepts. It considers the total number of paths between two concepts up to a user-specified maximum length. Authors from the same groups also introduced a co-occurence-based measure (CBM) [20], based on the same page counts as the NGD. Their measure does not require an estimation of the total number of Web pages. Note that this also means that the CBM of disjoint pairs of concepts are non-comparable. They argue for a combined connectivity/co-occurrence measure, since this would allow to find semantic relations between concepts that do not necessarily co-occur (with the connectivity-based measure),

[1] http://wordnet.princeton.edu/.
[2] http://www.nlm.nih.gov/mesh/.

while still emphasizing concept relations without the necessity of a strong connection in the semantic graph (with the CBM) [20].

A hybrid similarity metric for Linked Data was proposed in [16]. It considers the *information content* or *informativeness* of the features shared by two resources. The features of a resource are composed of the resources it links to, and the links themselves. In order for a feature to be shared, both the linked resource and the link type must be the same. Like the NSWD, the proposed metric requires an estimation of the number of concepts in the entire graph. It was specifically developed for a resource recommendation scenario and was preliminarily evaluated in such a scenario.

To sum up, most of the aforementioned instances of the NID are focused on comparison of purely textual, possibly ambiguous terms. Of those approaches mentioned above that do provide means to measure the distance between unambiguous concepts in a knowledge graph, none are based on the NID. Therefore, to the best of our knowledge, we provide the first NID-based (dis)similarity measure within the context of a knowledge graph.

4 Normalized Semantic Web Distance

The basic principle of the NSWD is to use the degree of co-occurrence of edges from and to two concepts in a knowledge graph to reflect their semantic dissimilarity. Instead of considering the human-understandable Web (accessed through a search engine), we consider a machine-understandable knowledge graph on the Semantic Web (accessed through a query client, e.g., a SPARQL endpoint).

The NSWD advances from possibly ambiguous natural language terms to unambiguous concepts, which are identified by URIs, as input. For example, when using the semantic dataset DBpedia.org, the island "Java" is uniquely identified by the URI `dbpedia:Java`[3], whereas the programming language "Java" is identified by `dbpedia:Java_(programming_language)`. We design the NSWD to result in a lower distance between `dbpedia:Java` and `dbpedia:Indonesia` than between `dbpedia:Java_(programming_language)` and `dbpedia:Indonesia`.

A knowledge graph consists of a set of nodes V and a set T of directed triples $t \in V \times P \times V$, where P is a set of predicates. A node represents a certain real-world object or concept and is identified via an URI. A triple $(u, p, v) \in T$ indicates a subject-predicate-object relation. The starting node u is the subject, v is the object and the predicate p carries some additional information regarding the exact nature of the relation between those nodes. For example, the triple *"dbpedia:Statue_of_Liberty dbpedia-owl:location dbpedia:New_York"* signifies the semantic relationship that the Statue of Liberty is located in New York.

To model the semantic relationship between two nodes in a knowledge graph, we make use of these triples. We consider the semantic relationship between two *subjects* to be stronger the more concepts they share a triple with. Additionally, the semantic relationship between two *objects* is considered stronger the more

[3] We choose the prefix `dbpedia:` for convenience, which resolves to http://dbpedia.org/resource/.

often they *occur* in a triple with the same subject, similar to a term occurring on a Web page. Nodes for which both of this is true, will have an even stronger relationship. We formalize these sets of linked nodes in Definition 3.

Definition 3. *We define the following sets of nodes* $V_\lambda \subseteq V$ *for* $\lambda \in \{in, out, all\}$ *in a knowledge graph* (V, T) *with respect to a node* $x \in V$*:*

$$V_{in}(x) := \{v \in V \mid (v, p, x) \in T\}$$
$$V_{out}(x) := \{v \in V \mid (x, p, v) \in T\}$$
$$V_{all}(x) := V_{in}(x) \cup V_{out}(x)$$

In other words, the set $V_{in}(x)$ comprises distinct nodes with at least one link – i.e., predicate – pointing to node x, whereas the set $V_{out}(x)$ contains all distinct nodes where node x points to. The set $V_{all}(x)$ is the union of all nodes that link to, or are linked to from node x (not necessarily with the same predicate). Based on these sets, we can now define three variations of frequency functions f_λ with respect to parameter $\lambda \in \{in, out, all\}$, to be used to calculate three variations of our proposed distance.

Definition 4. *For two nodes* x *and* y *in* V*, we define* f_λ *as:*

$$f_\lambda(x) := |V_\lambda(x)|$$
$$f_\lambda(x, y) := |V_\lambda(x) \cap V_\lambda(y)|$$

We then define the Normalized Semantic Web Distance *between two nodes* $x, y \in V$ *from a knowledge graph* (V, T) *as follows:*

$$\mathrm{NSWD}_\lambda(x, y) := \frac{\max\{\log f_\lambda(x), \log f_\lambda(y)\} - \log f_\lambda(x, y)}{\log N - \min\{\log f_\lambda(x), \log f_\lambda(y)\}}$$

As can be seen in the definition of the sets V_λ, the NSWD_λ makes use of all information available within the direct semantic context of the nodes in the corresponding knowledge graph. The parameter $\lambda \in \{in, out, all\}$ models the semantic context that is taken into account when determining the dissimilarity of two nodes. We investigate the question of which parameter λ performs best in practice in Sect. 8 and continue with and example of the NSWD_{in}.

Example 1. Consider the sub-graph of a graph with $|V| = 1000$, as depicted in Fig. 1. We want to calculate $\mathrm{NSWD}_{in}(x, y)$. Here, $V_{in}(x) = \{a, d, e, f, y\}$ is the set of concepts containing a link to x. Similarly, $V_{in}(y) = \{a, b, c, f\}$ is the set of concepts containing a link to y. Finally, the set of concepts with a link to both x and y is $V_{in}(x) \cap V_{in}(y) = \{a, f\}$. This means that $f(x) = |V_{in}(x)| = 5$, $f(y) = |V_{in}(y)| = 4$, and $f(x, y) = |V_{in}(x) \cap V_{in}(y)| = 2$. We therefore have $\mathrm{NSWD}_{in}(x, y) = \frac{log5 - log2}{log1000 - log4} \approx 0.16595$.

Note that the NSWD_{in} corresponds most closely to the original definition of the NWD (if we consider Web pages to be subjects and the search terms

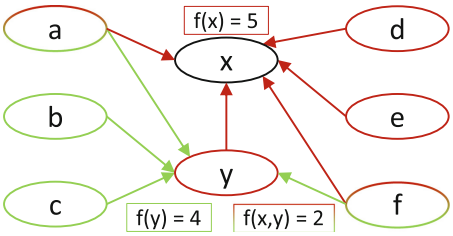

Fig. 1. Illustration of the values for the functions $f(x)$, $f(y)$, and $f(x, y)$ needed to calculate the NSWD_{in} in Example 1.

as part of its description). On the other hand, the NSWD_{out} corresponds more to the principles of the Jaccard Distance, since it relies on the shared outgoing links. Finally, the NSWD_{all} corresponds more closely to the principles used in the Wikipedia Link-based Measure [18], as it combines outgoing and incoming links, albeit using a different formula. Which of these definitions performs best in a real-world scenario can only be determined through empirical evaluation, which we describe in Sect. 8. First, however, we discuss some of the properties of the NSWD in Sect. 5, and how to implement it in Sect. 6.

5 Properties

In order to fully understand its potential, applicability, and implementation, we discuss a number of important properties of the NSWD, such as its mathematical properties, a number of special cases, and its minimum and maximum values.

Mathematical Properties. Although the NSWD is based on the NID, it is not a metric. For two arbitrary non-equal concepts x and y, $NSWD(x, x) = 0$ and $NSWD(x, y) \geq 0$ always hold, but not necessarily $NSWD(x, y) > 0$, since both concepts could be connected to exactly the same nodes. The NSWD also does not fulfill the triangle inequality. Two concepts can have very few links in common, making their NSWD high, while there could be a third term having many links in common with both concepts, making the NSWD between this third concept and each of the other two low. As highlighted by the authors of the NWD, which is also not a metric [4], this is not necessarily a drawback since human knowledge consists of many concepts which, intuitively, do not fulfill a triangle inequality. For example, while "Paris" and "White House" intuitively have a low semantic similarity, both have a high semantic similarity to "Capital".

Special Cases. A first special case occurs when $f_\lambda(x) = |V| = f_\lambda(y)$. In this case, $f_\lambda(x, y) = |V|$, so the numerator and denominator of $\text{NSWD}_\lambda(x, y)$ both become 0, and $\text{NSWD}_(x, y)$ does not exist. Note that in reality, this situation will never occur, since this would mean that both x and y are linked to by every element in the dataset, and thus do not contribute any useful information. However, for the sake of completeness, we define the NSWD to be 0 in this case.

A second special case occurs when $f_\lambda(x,y) = 0$. In case of $\lambda = in$, this means there are no concepts in the dataset that link to both x and y. In case of $\lambda = out$, it means that x and y do not link to any common concept. In these cases, $log f_\lambda(x,y)$ does not exist, and thus a value for the NSWD must be defined. For the Normalized Google Distance, the authors chose $\text{NGD}(x,y) = 1$ in case there were no search results that contained both x and y [5]. However, that leads to a – in our view – very counter-intuitive situation where it is possible for two terms that have a low, but strictly positive number of search results for both terms, to have a higher NGD than two terms that do not. In fact, the first special case adds to this argument, since the distance is 1, whereas there is clearly some relationship between x and y (however abstract it may be). A much more elegant solution is to determine the upper bound NSWD_{max} for NSWD_λ, and define $\text{NSWD}_\lambda(x,y) = \text{NSWD}_{max}$ in the case that $f_\lambda(x,y) = 0$.

A final special case occurs if $f_\lambda(x) = 0$ (or $f_\lambda(y) = 0$), in which case the $log f_\lambda(x)$ (or $log f_\lambda(y)$) does not exist. In this case, $f_\lambda(x,y)$ will automatically equal 0 as well, bringing us back to the second special case. Therefore, we will define $\text{NSWD}_\lambda(x,y) = \text{NSWD}_{max}$ in this case as well.

Maximum Value. With the values for the special cases as defined above in mind, we can rely on our knowledge of the behavior of $f_\lambda(x,y)$ with respect to $f_\lambda(x)$ and $f_\lambda(y)$ to calculate NSWD_{max}. It can be shown that

$$\forall x, y \in V : f_\lambda(x) + f_\lambda(y) - |V| \leq f_\lambda(x,y) \leq |V| - 1$$

Keeping these constraints in mind, we can iterate over all possible values of $\text{NSWD}_\lambda(x,y)$ for a given cardinality of V. This way, we determined that:

$$\forall x, y \in V : \text{NSWD}_{max} = \frac{log(\lfloor \frac{|V|}{2} \rfloor + 1)}{log\,|V| - log\lceil \frac{|V|}{2} \rceil}$$

For example, considering the same dataset V as in Example 1, the upper bound for $\text{NSWD}_\lambda(x,y)$ for any concepts x and y in V is $\text{NSWD}_{max} = \frac{log501}{log1000 - log500} \approx 8.9686$. We can now use this value to determine the NSWD in the special case where $f_\lambda(x,y) = 0$, as well as to calculate a similarity score normalized between 0 and 1, as further explained in Sect. 7.

6 Implementation

The NSWD is not bound to one specific technology for its implementation. Any method that can calculate the frequency functions as described in Definition 4, as well as a count of the number of nodes in a knowledge graph is suitable. This makes the NSWD a very flexible measure, allowing it to be implemented using the most suitable technology for any specific use case. Additionally, it can be tailored to any knowledge domain by using a specialized dataset. However, typically, a knowledge graph on the Semantic Web is modeled in RDF, and it is made accessible through SPARQL. For any knowledge graph with a SPARQL

endpoint, the following query can be used to calculate $f_\lambda(x)$ and $f_\lambda(x, y)$: `SELECT (COUNT(DISTINCT ?a) AS ?f) WHERE`, with the matching `WHERE`-clause from Table 1, and x and y the URIs of two concepts in the knowledge graph.

Table 1. `WHERE`-clauses for SPARQL queries to calculate the NSWD.

Frequency function	`WHERE`-clause
$f_{in}(x)$	`{ ?a ?p <x> }`
$f_{in}(x, y)$	`{ ?a ?p1 <x> . ?a ?p2 <y> }`
$f_{out}(x)$	`{ <x> ?p ?a }`
$f_{out}(x, y)$	`{ <x> ?p1 ?a . <y> ?p2 ?a }`
$f_{all}(x)$	`{{ ?a ?p1 <x> } UNION { <x> ?p2 ?a }}`
$f_{all}(x, y)$	`{{?a ?p1 <x>} UNION {<x> ?p2 ?a} . {?a ?p3 <y>} UNION {<y> ?p4 ?a}}`

To calculate the total number of distinct concepts in the knowledge graph, a `COUNT DISTINCT` query could be used. However, due to the `DISTINCT` constraint, such a query – while conceptually straightforward – does pose a problem for larger knowledge graphs, as most SPARQL endpoints are not optimized for this kind of count. In these cases, it is advised to store a cached result for an earlier execution of this query, a value from the dataset's metadata (if available), or an estimate of the number of nodes in the knowledge graph, and use this for the calculations. Optimizing for these types of queries would be a must to guarantee the usability of the NSWD on a large scale.

7 Calculating Similarity

The NSWD is a *distance*, meaning that the more semantically related two concepts are, the smaller their distance is. However, in many cases the opposite is desired, i.e. when *similarity* must be measured – e.g., in the Miller-Charles benchmark [17], which we use for our evaluation. If the NSWD were normally distributed in the range $[0, \text{NSWD}_{max}]$, we could just scale it linearly using NSWD_{max} and substract it from 1. However, the values that occur most frequently are in the $[0, 1]$ range. These are also the distance values that are most interesting in practical scenarios, such as recommendation systems. Scaling these values linearly using the NSWD_{max} would lead to a situation where the majority of distances would be in the range of $[0, \frac{1}{\text{NSWD}_{max}}]$, which is not very useful. Keeping this in mind, we define the NSWD-based similarity Sim_{NSWD} as follows:

Definition 5.

$$Sim_{NSWD_\lambda}(x, y) := \begin{cases} 1 - d(x, y) \times (1 - c), & \text{if } d(x, y) \in [0, 1] \\ (1 - \frac{d(x,y)}{NSWD_{max}}) \times c, & \text{if } d(x, y) \in]1, NSWD_{max}] \end{cases}$$

$$\text{with } d(x, y) = NSWD_\lambda(x, y) \text{ and } c = \frac{1}{NSWD_{max}}$$

This way, the most semantically significant distances – those between 0 and 1 – get mapped to the similarity range $[c, 1]$ with minimal scaling, and the distances higher than 1 get mapped to the similarity range $[0, c[$ with significant scaling. Note that if NSWD_{max} is accurately calculated, $Sim_{\text{NSWD}_\lambda}(x, y)$ is normalized between 0 and 1.

8 Evaluation

In this section, we first compare the NSWD-based similarity measures defined in Sect. 7 to human judgment of similarity, using a standard set of term-pairs. Next, we verify whether the NSWD and its variants accomplish what they were designed to do: increase semantic awareness w.r.t traditional distance measures.

To evaluate our approach, we chose to use the Miller-Charles dataset [17], which consists of 30 term-pairs that were judged for similarity by 38 people. While using lexical terms for evaluation of a distance in knowledge graphs is not ideal, and this is a relatively small dataset, it offers an insight to how humans judge the similarity between these terms, and more importantly, it gives us a number of related approaches for direct comparison, as it is very commonly used in this field. Therefore, the Miller-Charles benchmark provides the best starting point for external validation compared to established approaches. However, before we can use it, a disambiguation strategy has to be decided upon, as many of the terms are highly ambiguous, in the sense that they can correspond to more than one resource URI in the knowledge graph. To choose one to use for the NSWD calculation, we used three disambiguation strategies:

Manual: manually pick a disambiguated resource URI, or suggest an alternative URI (human judgment);
Count-Based: use the resource URI with the highest V_{in}, V_{out} or V_{all} (depending on whether the NSWD, NSWD_{out} or NSWD_{all} is calculated, respectively);
Similarity-Based: use the resource URI leading to the smallest distance (only possible in the context of a pairwise comparison);

Note that it is possible that the correct disambiguation cannot be determined due to the non-completeness of the dataset. In our evaluation, we calculated the distances using all aforementioned disambiguation strategies, to see which leads to the best results. For each of the 30 term-pairs, our evaluation process consists of the steps below.

1. Both terms are disambiguated, using the manual and automatic approaches. This results in 3 URI disambiguation options for each term: (a) manually selected, (b) based on the highest link-count, and (c) based on the highest similarity with the other term.
2. For each of the three URI disambiguation options, the NSWD_{in}, NSWD_{out}, and NSWD_{all} are calculated.
3. The above results in 9 distances (three for each variant of the NSWD), which are converted NSWD-based similarities, as defined in Definition 5.

These steps result in 9 similarity assessments for each of the 30 term-pairs, each value calculated with a different combination of disambiguation option and NSWD variant. These values are compared to the human-assessed scores from the Milled-Charles dataset by calculating the Pearson correlation coefficient.

As a baseline, we added a similarity score based on the Normalized Bing Distance [10] to the evaluation results. This NBD-based similarity was calculated as $1 - \mathrm{NWD}(x, y)$, with NWD(x, y) calculated as specified in Definition 2, using the Microsoft Bing Search API[4] as a search engine.

We calculated the Pearson correlation coefficient between the Miller-Charles scores and the NBD baseline, as well as the three NSWD variants. Each NSWD-based similarity measure was tested with three disambiguation strategies: manual (M), count-based (C), or similarity-based (S), using two widely used knowledge graphs: Freebase and DBpedia. The results are shown in Table 2, along with the reported correlations on the same benchmark for the Wikipedia Link-based Measure* [18], and Jaccard similarity as calculated in [13]. Higher correlation indicates a stronger positive relationship between the human-assessed scores and calculated similarities. To enable reproducibility of the results, we provide online access to the JSON files generated by our evaluation software, including all disambiguated URIs, Miller-Charles scores, and similarity scores. The file for Freebase can be accessed at http://semweb.mmlab.be/nswd/evaluation/mc30_results_freebase.json, the file for DBpedia at http://semweb.mmlab.be/nswd/evaluation/mc30_results_dbpedia.json, and the file for Bing at http://semweb.mmlab.be/nswd/evaluation/mc30_results_bing.json.

Table 2. Pearson correlation coefficient on the Miller-Charles benchmark for the NSWD similarity variants on the Freebase and DBpedia knowledge graphs, as well as the Normalized Bing Distance, Wikipedia Link-based Measure, and Jaccard similarity.

NBD	**0.23**								
	$Sim_{\mathrm{NSWD}_{in}}$			$Sim_{\mathrm{NSWD}_{out}}$			$Sim_{\mathrm{NSWD}_{all}}$		
	M	C	S	M	C	S	M	C	S
Freebase	0.42	0.25	0.29	0.57	0.43	0.57	0.55	0.24	**0.58**
DBPedia	0.60	0.44	0.55	0.56	**0.69**	0.62	0.66	0.58	0.68
WLM[a]	**0.70**								
Jaccard[a]	**0.882**								

[a]Using a different disambiguation strategy

Note that for all distance and disambiguation options, the NSWD-based similarities achieved a higher correlation than the NBD-based similarity at the time of writing, with a maximum of 0.58 for Freebase and 0.69 for DBpedia. There is no consistent trend in which disambiguation strategy performed best. Overall,

[4] https://datamarket.azure.com/dataset/bing/search.

the NSWD_{all} seemed to perform best, taking most of the semantic context of a node into account. None of the NSWD variants was able to perform better than the reported results of the WLM and Jaccard similarity. However, note that for these reported results, a different disambiguation strategy was applied. For the WLM as reported in [18], the disambiguation of the Miller-Charles terms was performed using a weighted combination of commonness, relatedness, and occurrence together in a sentence. However, the authors of [18] did not disclose the exact weighting scheme they used, nor the disambiguated terms. Do note that commonness and relatedness of the terms in a term-pair are factors that we also consider, by applying the disambiguation strategies using the highest link-count and highest similarity, respectively. Therefore, we can safely assume that the reported correlation of 0.70 is useful to compare with our results. In case of the Jaccard similarity as calculated in [13], disambiguation was left ad-hoc to a search engine, which makes it impossible for us to reproduce.

While the aforementioned results show the external validity of our approach, they do not highlight semantic awareness – the primary aspect it was designed for. To highlight this aspect of the NSWD, we created a small additional evaluation set of our own, consisting of 10 pairs of concept-pairs with the same plain-text representation, yet with the first concept in both concept-pairs remaining the same, and the second concept disambiguated to a more divergent meaning. While this is a limited set of concepts, the results clearly illustrate that the NSWD and its variants are capable of recognizing these differences in semantics, and assign smaller distances to concepts that are close in semantics than to concepts whose semantics diverge.

The evaluation set for the DBpedia knowledge graph is shown in Table 3, along with the results. We observe that in all 10 cases, all three variants of the NSWD maintain a semantically correct ordering w.r.t. the distance. Note that the impact of the lower link-density of DBpedia is seen here as well, with many distances being equal to 21.2, which is the NSWD_{max} for DBpedia. This means that in all these cases there were no common links, and thus $f(x, y)$ was zero.

To use the same evaluation set for the Freebase graph, we mapped every DBpedia resource from Table 3 to a Freebase resource using the `owl:sameAs` link to Freebase included in the description of the resource[5]. As shown in Table 4, the NSWD and its variants also maintain a semantically correct ordering w.r.t. the distance when applied to Freebase. We observed one small discrepancy for the resource-pairs `:Automobile-:Bus` and `:Automobile-:Bus_(computing)`, where the ordering is (very slightly) reversed for the NSWD.

The quality of the knowledge graph greatly affects the performance and applicability of the NSWD. For example, during the disambiguation of the Miller-Charles dataset, we found that DBpedia often lacks a simple description of various concepts. For example, the concepts "journey" and "voyage" – resulting in the resources `dbpedia:Journey` and `dbpedia:Voyage`, repectively – both link to many disambiguation options, but none of these options capture the most straightforward meaning of the concepts. When inspecting the

[5] The full mapping is available at http://semweb.mmlab.be/nswd/evaluation/resourcemapping_dbpedia_freebase.ttl.

Table 3. Results illustrating the semantic awareness of the NSWD variants using the DBpedia graph. The prefix : resolves to http://dbpedia.org/resource/. The 'order'-column indicates whether $d(c1, c2) < d(c1, c3)$, which is desired in these cases.

concept 1 (c1)	concept 2 (c2)	concept 1 (c3)	d = NSWD			d = NSWD$_{out}$			d = NSWD$_{all}$		
			$d(c1,c2)$	$d(c1,c3)$	order	$d(c1,c2)$	$d(c1,c3)$	order	$d(c1,c2)$	$d(c1,c3)$	order
:Pear	:Apple	:Apple_Inc.	0.27	21.2	√	0.16	21.2	√	0.21	21.2	√
:Trout	:Bass_(fish)	:Bass_guitar	21.2	21.2	-	0.31	21.2	√	0.33	21.2	√
:Cat	:Jaguar	:Jaguar_Cars	21.2	21.2	-	0.13	0.40	√	0.20	0.50	√
:Cat	:Mouse	:Mouse_(computing)	0.41	21.2	√	0.13	21.2	√	0.21	21.2	√
:Automobile	:Bus	:Bus_(computing)	0.33	21.2	√	21.2	21.2	√	0.34	21.2	√
:Indonesia	:Java	:Java_(programming_language)	0.39	21.2	√	0.24	0.49	√	0.39	1.01	√
:Lion	:Tiger	:Tiger_(Danish_store)	0.39	21.2	√	0.14	21.2	√	0.23	21.2	√
:Musical_theatre	:Broadway_(play)	:Broadway_(Manhattan)	21.2	21.2	-	0.34	0.37	√	0.51	0.58	√
:Bird	:Crane_(bird)	:Crane_(machine)	0.55	21.2	√	0.26	21.2	√	0.58	21.2	√
:Bass_guitar	:String_(music)	:String_(physics)	0.53	21.2	√	0.33	21.2	√	0.56	21.2	√

Table 4. Results illustrating the semantic awareness of the NSWD variants using the Freebase graph. For clarity, the DBpedia concept names are shown instead of the Freebase hash codes. For each DBpedia resource, the actual Freebase URI used in the calculations is the object of the `owl:sameAs` relation to the corresponding Freebase resource.

concept 1 (c1)	concept 2 (c2)	concept 1 (c3)	d = NSWD			d = NSWD$_{out}$			d = NSWD$_{all}$		
			$d(c1,c2)$	$d(c1,c3)$	order	$d(c1,c2)$	$d(c1,c3)$	order	$d(c1,c2)$	$d(c1,c3)$	order
:Pear	:Apple	:Apple_Inc.	0.13	0.37	√	0.17	0.46	√	0.17	0.44	√
:Trout	:Bass_(fish)	:Bass_guitar	0.15	0.57	√	0.32	0.67	√	0.32	0.68	√
:Cat	:Jaguar	:Jaguar_Cars	0.25	0.35	√	0.31	0.46	√	0.32	0.47	√
:Cat	:Mouse	:Mouse_(computing)	0.23	0.33	√	0.31	0.46	√	0.31	0.47	√
:Automobile	:Bus	:Bus_(computing)	0.34	0.32	×	0.38	0.46	√	0.38	0.46	√
:Indonesia	:Java	:Java_(programming_language)	0.30	0.58	√	0.33	0.53	√	0.36	0.60	√
:Lion	:Tiger	:Tiger_(Danish_store)	0.11	0.28	√	0.22	0.42	√	0.21	0.43	√
:Musical_theatre	:Broadway_(play)	:Broadway_(Manhattan)	0.42	0.43	√	0.42	0.46	√	0.44	0.50	√
:Bird	:Crane_(bird)	:Crane_(machine)	0.22	0.33	√	0.33	0.51	√	0.33	0.52	√
:Bass_guitar	:String_(music)	:String_(physics)	0.46	0.57	√	0.46	0.73	√	0.46	0.73	√

corresponding human-understandable Wikipedia pages, it becomes clear that both "journey" and "voyage" are supposed to be disambiguated to the concept "travel", with resource URI `dbpedia:Travel`. Unfortunately, these links are not currently included in DBpedia. As a result, automatic disambiguation methods (such as the count-based and similarity-based disambiguation) that only follow links included in the knowledge graph will never find the correct result, leaving manual disambiguation by a human as the only correct option in these cases. In a number of other cases, no resource exists to represent a concept, as was the case with the terms "lad" and "madhouse". The lower connectivity between concepts in DBpedia also resulted in many of the distances defaulting to NSWD$_{max}$ during the evaluation. Freebase was found to be richer in this regard, as we found less zero-scores, and smaller variances in the similarities than in the DBpedia results. Concepts in Freebase were missing for fewer terms than in DBpedia, and there were less cases where two terms in a term-pair corresponded to the same URI. Still, terms such as "lad" and "madhouse" have no direct equivalent on Freebase.

9 Conclusion and Future Work

In this paper, we investigated the Normalized Semantic Web Distance: a semantically aware adaptation of the Normalized Information Distance, relying on links in a knowledge graph. We described three variations, taking into account incoming links, outgoing links, or both. We discussed the properties and special cases, and we proposed a conversion of the NSWD to measure similarity, using a customized normalization scheme. We extensively evaluated our approach, ensuring external validity by choosing an established benchmark: the Miller-Charles dataset of 30 human-assessed term-pairs. When applied to the Freebase knowledge graph, the NSWD and its variants exhibit a correlation of up to 0.58 with human similarity assessments, and when applied to DBpedia, the correlation was even higher at 0.69, albeit with less fine-grained similarity scores due to DBpedia's smaller size. We also verified that the NSWD maintains semantic awareness when confronted with ambiguous concept-pairs.

In future work, we will further illustrate the merit of the NSWD variants by applying them on more domain-specific knowledge graphs. We suspect that if the domain knowledge of the graph is high, the NSWD variants should be aware of these semantics and perform better than traditional approaches. Additionally, we aim to gather a larger set of concepts with an ambiguous plain-text representation as used to illustrate the semantic awareness of the NSWD variants, using a more systematic approach, supported by human assessments.

Acknowledgments. The research activities in this paper were funded by Ghent University, iMinds (by the Flemish Government), IWT Flanders, FWO-Flanders, the European Union, and RWTH Aachen University.

References

1. Beecks, C., Uysal, M.S., Seidl, T.: Signature quadratic form distance. In: Proceedings of the ACM International Conference on Image and Video Retrieval, pp. 438–445. ACM (2010)
2. Ceccarelli, D., Lucchese, C., Orlando, S., Perego, R., Trani, S.: Learning relatedness measures for entity linking. In: 22nd ACM International Conference on Information and Knowledge Management, pp. 139–148. ACM (2013)
3. Cilibrasi, R., Vitanyi, P.: Clustering by compression. IEEE Trans. Inf. Theory **51**(4), 1523–1545 (2005)
4. Cilibrasi, R., Vitányi, P.M.B.: Normalized web distance and word similarity. CoRR abs/0905.4039 (2009)
5. Cilibrasi, R.L., Vitanyi, P.M.B.: The Google similarity distance. IEEE Trans. Knowl. Data Eng. **19**(3), 370–383 (2007)
6. De Nies, T., Beecks, C., De Neve, W., Seidl, T., Mannens, E., Van de Walle, R.: Towards named-entity-based similarity measures: challenges and opportunities. In: ESAIR, pp. 9–11. ACM(2014)

7. De Nies, T., Beecks, C., Godin, F., De Neve, W., Stepien, G., Arndt, D., De Vocht, L., Verborgh, R., Seidl, T., Mannens, E., Van de Walle, R.: A distance-based approach for semantic dissimilarity in knowledge graphs. In: Proceedings of the 10th International Conference on Semantic Computing (ICSC, TBP). IEEE (2016)
8. De Vocht, L., Coppens, S., Verborgh, R., Vander Sande, M., Mannens, E., Van de Walle, R.: Discovering meaningful connections between resources in the web of data. In: Proceedings of the 6th Workshop on Linked Data on the Web (2013)
9. Eskevich, M., Jones, G.J., Aly, R., Ordelman, R., Chen, S., Nadeem, D., Guinaudeau, C., Gravier, G., Sébillot, P., De Nies, T., et al.: Multimedia information seeking through search and hyperlinking. In: Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval, pp. 287–294 (2013)
10. Godin, F., De Nies, T., Beecks, C., De Vocht, L., De Neve, W., Mannens, E., Seidl, T., Van de Walle, R.: The normalized Freebase distance. In: 11th Extended Semantic Web Conference (2014)
11. Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E.G., Milios, E.: Information retrieval by semantic similarity. Int. J. Semant. Web Inf. Syst. (IJSWIS) **2**(3), 55–73 (2006)
12. Jacobs, I., Walsh, N., et al. (eds.): Architecture of the World Wide Web, Volume One. W3C Recommendation 15 December 2004
13. Kulkarni, S., Caragea, D.: Computation of the semantic relatedness between words using concept clouds. In: Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, pp. 183–188 (2009)
14. Li, M.: An Introduction to Kolmogorov Complexity and Its Applications. Springer, Heidelberg (1997)
15. Li, M., Chen, X., Li, X., Ma, B., Vitanyi, P.: The similarity metric. IEEE Trans. Inf. Theory **50**(12), 3250–3264 (2004)
16. Meymandpour, R., Davis, J.G.: Recommendations using linked data. In: Proceedings of the 5th Ph.D. workshop on Information and knowledge, pp. 75–82 (2012)
17. Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. Lang. Cogn. Process. **6**(1), 1–28 (1991)
18. Milne, D., Witten, I.: An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In: Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy, pp. 25–30, Chicago, USA (2008)
19. Moore, J.L., Steinke, F., Tresp, V.: A novel metric for information retrieval in semantic networks. In: García-Castro, R., Fensel, D., Antoniou, G. (eds.) ESWC 2011. LNCS, vol. 7117, pp. 65–79. Springer, Heidelberg (2012)
20. Nunes, B.P., Dietze, S., Casanova, M.A., Kawase, R., Fetahu, B., Nejdl, W.: Combining a co-occurrence-based and a semantic measure for entity linking. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) The Semantic Web: Semantics and Big Data. LNCS, vol. 7882, pp. 548–562. Springer, Heidelberg (2013)
21. Nunes, B.P., Herrera, J., Taibi, D., Lopes, G.R., Casanova, M.A., Dietze, S.: SCS connector-quantifying and visualising semantic connections between entity pairs. In: Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A. (eds.) The Semantic Web: ESWC 2014 Satellite Events. LNCS, vol. 8798, pp. 461–466. Springer, Heidelberg (2014)
22. Passant, A.: Measuring semantic distance on linking data and using it for resources recommendations. In: AAAI Spring Symposium: Linked Data Meets Artificial Intelligence (2010)

23. Sánchez, D., Batet, M., Isern, D., Valls, A.: Ontology-based semantic similarity: a new feature-based approach. Expert Syst. Appl. **39**(9), 7718–7728 (2012)
24. Schwartz, B.: Google removes the + search command (2011). http://searchengineland.com/google-sunsets-search-operator-98189
25. Zuo, Z., Huang, H.H., Kawagoe, K.: Similarity search of human behavior processes using extended linked data semantic distance. In: 25th International Workshop on Database and Expert Systems Applications (DEXA), pp. 178–182. IEEE (2014)