Correlation of Ontology-Based Semantic Similarity and Human Judgement for a Domain Specific Fashion Ontology

Edgar Kalkowski(✉) and Bernhard Sick

University of Kassel, Kassel, Germany
{kalkowski,bsick}@uni-kassel.de

Abstract. Evaluation of semantic similarity is difficult because semantic similarity values are highly subjective. There are several approaches that compare automatically computed similarities with values assigned by humans for general purpose terms and ontologies that contain general purpose terms. However, ontologies should be as domain specific as possible to capture the maximal amount of semantic knowledge about a domain. To evaluate the semantic knowledge captured by a custom fashion ontology we conducted a survey and crowdsourced similarity values for fashion terms. In this article we compare the manually assigned similarities to those computed automatically with several ontology-based similarity measures. We show that our proposed feature-based measure achieves the highest correlation with human judgement and give some insight into why this kind of similarity measure most resembles human similarity assessments. To evaluate the influence of the ontology on similarities we compare the results achieved with our fashion ontology to similarity values computed using a fragment of DBpedia.

Keywords: Feature based similarity · Semantic similarity · Fashion ontology

1 Introduction

Many applications of the semantic web deal with ontologies and use them to assess the semantic similarity of terms. However, since the semantic similarity of terms is highly subjective it is difficult to evaluate whether or not computed similarity values actually make sense. One approach is to compare automatically computed similarities to values assigned by human experts. For example, in [1] 353 and in [2] 3000 word pairs were manually assigned with similarity values based on human judgement. In both cases general purpose words were used and WordNet [3] was used to automatically compute similarities for comparison.

Our application is concerned with search engine marketing where much historical data is aggregated over time for all search keywords entered into the search engine and the advertisement displayed for them. However, since there can be thousands [4] or millions [5,6] of keywords, data for many low traffic

A. Bozzon et al. (Eds.): ICWE 2016, LNCS 9671, pp. 207–224, 2016.
DOI: 10.1007/978-3-319-38791-8_12

keywords is sparse, if existent at all. Since those keywords are rather specific, they are interesting for advertisement, because it is reasonable to assume that a customer is more inclined to buy an advertised product if it fits to their very specific search request.

To derive forecasts for low traffic keywords we propose to aggregate data from similar keywords. These aggregated data can then be used to train machine learning models like ARMA models [7–9], support vector regression [10–12], or generative models such as those presented in [13,14]. To compute similarities between keywords we propose to use ontology-based similarity measures. Because the terms that occur in search engine marketing are highly specific to the domain for which ads are placed we created our own custom ontology that covers those domain specific terms.

The key contribution of this article is an evaluation of the similarity values computed based on our custom ontology. Since the terms are highly domain specific the results of [1,2] cannot be used for comparison since very few terms from our search engine marketing domain occur in these samples. Instead, we gathered similarity values assigned by humans for comparison by conducting our own survey among 183 participants. In this article we evaluate the results of the survey and compare the similarities assigned by humans with similarity assessments computed automatically by similarity measures based on our custom ontology. In addition to the custom ontology we also used a subset of the DBpedia ontology [15] extracted from Wikipedia for the evaluation.

The remainder of this article is structured as follows: In Sect. 2 we give a brief overview of related work. The ontologies we used in our evaluation are described in Sect. 3 and the similarity measures based on those ontologies are defined in Sect. 4. Section 5 is concerned with the survey we conducted and in Sect. 6 we discuss our findings and compare the similarity values obtained with the survey with those computed by our similarity measures. Finally, in Sect. 7 we summarize our findings and give a brief outlook of further research.

2 Related Work

This articles touches on two particular topics for both of which some related work will be presented in this section. Firstly, in this article several graph based similarity measures are used. Secondly, this article is concerned with evaluating the semantics captured by a custom ontology by comparing automatically computed similarity values to those assigned by humans.

There are several approaches to computing similarities based on graphs or ontologies. The article [16] gives an overview of ontology-based similarity measures and categorizes measures as being either *edge counting*, *feature-based*, or making use of *information content*. In case of edge counting measures [17–20] the similarity assessment of two nodes in an ontology is based on the number of hops on a path between the two nodes for edge counting measures. Feature-based similarity measures [21,22] assess the properties of nodes in the ontology, e.g. their taxonomical neighborhood. Finally, measures making use of information content [23–27] are based on a big text corpus and assess the similarity of nodes in an ontology by evaluating the frequencies and positions in which terms of the nodes occur in the text corpus.

In this article we use two edge counting approaches and one feature-based method to compute the similarity of concepts in ontologies. Measures based on information content are not feasible in our application due to the lack of a suitable domain specific text corpus.

As mentioned for example in [28] it is difficult to evaluate whether or not semantic similarity values are reasonable since the semantic similarity of two terms is highly subjective. However, there have been several approaches [1,2,29,30] in which automatically computed similarities have been compared to those assigned by humans. All of these data sets use general purpose terms and only few general purpose terms occur in our domain specific ontology. Thus, in order to evaluate how well our ontology captures semantic information about our application domain we performed a survey to create a new data set with manually assigned similarity values for term pairs taken from the fashion domain.

3 Ontologies

In this section we briefly describe our custom fashion ontology and the fragment of DBpedia we use for similarity measurements.

There already exist many general purpose ontologies like DBpedia [15] or WordNet [3] which are also available in German. However, our application contains many domain specific terms, especially fashion brands and categories, which are not represented in general purpose ontologies. Thus, we created our own ontology which tries to incorporate as many relevant terms as possible by analyzing the website of the online shop doing the advertising, in our case a shop mainly concerned with fashion items.

The main hierarchical structure of our ontology is a tree of 769 fashion categories. Orthogonal to the categories we have several secondary flat hierarchies that describe properties of the fashion items present in each category, e.g., used materials, color, size, etc. The biggest secondary hierarchy is that of fashion brands which contains 1749 entries, even more than we have fashion categories. Table 1 summarizes all concepts in our ontology and additionally states the number of instances of each concept and the number of connections from instances of each concept to other nodes in the ontology.

There are several options to store an ontology. We decided to use Neo4J [31] since this database explicitly supports graph storage and natively supports typical graph queries such as paths of arbitrary lengths between two nodes which is convenient for edge counting approaches.

Our fashion ontology is the first ontology used in this paper. To analyze the influence of the ontology on similarity measurements we compare the results obtained with our fashion ontology to those computed with a fragment of the DBpedia. The nodes of this ontology represent Wikipedia pages and the relationships between nodes represent links between Wikipedia pages. Including further

Table 1. Concepts and relationships in our fashion ontology. For each concept the number of instances in the ontology is stated and the number of connections from those instances to other nodes in the ontology. Where appropriate the concepts were translated from German.

Concept	Instances	Connections	Concept	Instances	Connections
Brand	1749	90748	Pattern	7	1118
Category	769	1520	Clasp	6	872
Size	522	30932	HeelHeight	5	378
Technology	24	1204	ShaftHeight	5	220
OuterMaterial	21	6356	DEN	5	24
Color	18	17162	ShoeTip	4	542
InnerLining	17	1468	ShaftWidth	3	60
Length	15	362	TrouserHeight	3	136
Collar	12	580	Collection	2	1316
HeelForm	10	584			

Wikipedia content such as the info boxes or article abstracts into the ontology is not feasible since the resulting ontology gets too big to be stored in the Neo4J database and query times become unreasonably high and memory intensive. The DBpedia fragment we use consists of 7 149 395 instances of the "resource" concept and 112 453 671 connections between resources each of which represents a link between the two corresponding Wikipedia pages.

The search engine marketing application is concerned with 236 837 keywords which consist of 26 324 unique terms. With our fashion ontology for 201 789 of the keywords at least one matching concept in the ontology can be found. For the DBpedia this number is a little higher since we find at least one concept for 224 102 keywords. When looking directly at the terms, the difference gets even higher: We find at least one concept for only 3 553 terms using our fashion ontology while 15 575 terms can be mapped to at least one concept in the DBpedia ontology. Although the coverage of both keywords and terms is greater in case of the DBpedia ontology we show in Sect. 6 that our fashion ontology better captures the semantics of the application domain in the sense that similarities computed with the fashion ontology have a higher correlation to human judgement.

After we described the used ontologies in this section the following section is concerned with the similarity measures which make use of the semantic information stored in the ontologies to compute the similarities of sets of terms.

4 Similarity Measures

This section presents the ontology-based similarity measures used in this article.

The goal of our measures is to compute the similarity of two sets of terms $\mathcal{T}_1$ and $\mathcal{T}_2$ whose terms belong to one of two keywords that shall be compared.

The comparison is executed in three stages: First, we compare the terms themselves and factor in how many terms the two sets have in common. Secondly, the terms are mapped to concepts in the used ontology and a second factor considers how many concepts are common in both sets. Lastly, we compare the remaining concepts using the structure of the ontology in different ways.

The first factor that considers the common terms is denoted with τ and is computed as

$$\tau = \begin{cases} 0, & \mathcal{T}_1 = \mathcal{T}_2 = \emptyset \\ \frac{|\mathcal{T}_1 \cap \mathcal{T}_2|}{|\mathcal{T}_1 \cup \mathcal{T}_2|}, & \text{otherwise} \end{cases} \tag{1}$$

where $|\cdot|$ denotes the cardinality of a set and $\mathcal{T}_1, \mathcal{T}_2$ are two sets of terms whose similarity shall be computed. If both sets have no terms in common or both sets are empty τ becomes 0, if both sets contain exactly the same terms τ is 1.

The remaining terms in each set which are not contained in the other set are mapped to concepts in the ontology. How exactly this is done depends on the used ontology. For our custom fashion ontology each concept node just contains one or more terms as a description. In this case we perform a case insensitive match and check if a term is contained as a substring in any of the terms describing a concept. For the DBpedia fragment each concept is described by a unique URI which always begins with http://de.dbpedia.org/resource. Here, the first common part of the URI is stripped to prevent terms like "http" or "resource" matching all nodes. The remainder of the URI is again searched in a case insensitive manner to check if a term is contained as a substring.

By means of this matching we get a set $\mathcal{C}'_1$ of concepts in the ontology for $\mathcal{T}_1 \backslash (\mathcal{T}_1 \cap \mathcal{T}_2)$ and a set $\mathcal{C}'_2$ of concepts and for $\mathcal{T}_2 \backslash (\mathcal{T}_1 \cap \mathcal{T}_2)$. The second factor then considers the concepts both sets have in common and is computed as

$$\zeta_{\mathcal{O}} = \begin{cases} 0, & \mathcal{C}'_1 = \mathcal{C}'_2 = \emptyset \\ \frac{|\mathcal{C}'_1 \cap \mathcal{C}'_2|}{|\mathcal{C}'_1 \cup \mathcal{C}'_2|}, & \text{otherwise} \end{cases} . \tag{2}$$

Similar to τ the factor $\zeta_{\mathcal{O}}$ is 1 if both sets of concepts are identical and becomes 0 if the two sets of concepts are disjoint. The index $\mathcal{O}$ indicates that in contrast to τ the factor $\zeta_{\mathcal{O}}$ depends on the used ontology $\mathcal{O}$.

For the final comparison of remaining concepts we remove common concepts from each of the two sets and get $\mathcal{C}_1 = \mathcal{C}'_1 \backslash (\mathcal{C}'_1 \cap \mathcal{C}'_2)$ and $\mathcal{C}_2 = \mathcal{C}'_2 \backslash (\mathcal{C}'_1 \cap \mathcal{C}'_2)$. The remaining concepts in the two disjoint sets $\mathcal{C}_1$ and $\mathcal{C}_2$ are now compared making use of the ontology in different ways. Overall, the similarity $\mathrm{sim}_{\mathcal{O},s_{\mathcal{O}}}(\mathcal{T}_1, \mathcal{T}_2)$ of the two sets of terms is computed as

$$\mathrm{sim}_{\mathcal{O},s_{\mathcal{O}}}(\mathcal{T}_1, \mathcal{T}_2) = \begin{cases} \tau + (1-\tau)\zeta_{\mathcal{O}}, & \mathcal{C}_1 = \emptyset \vee \mathcal{C}_2 = \emptyset \\ \tau + (1-\tau)(\zeta_{\mathcal{O}} + (1-\zeta_{\mathcal{O}}) s_{\mathcal{O}}(\mathcal{C}_1, \mathcal{C}_2)), & \text{otherwise} \end{cases} \tag{3}$$

where $s_{\mathcal{O}}$ is a similarity measure that assesses the similarity of the remaining disjoint sets of concepts $\mathcal{C}_1$ and $\mathcal{C}_2$.

Our main similarity measure is feature-based and considers how many neighbors each pair of concepts has in common and is thus called *direct neighbors*. This idea is inspired by the Google similarity distance [24] which was also adapted to Wikipedia [27]. We already proposed this measure in [32] and compared it to two edge-counting approaches. The idea of the first measure is that two concepts are very similar in case they have many common neighbors and they are very different in case they have few or no neighbors in common. In order to compute the neighborhood of nodes in the ontology graph let $n_{\mathcal{O}} : \mathcal{C} \rightarrow \mathcal{P}(\mathcal{C})$ be a function that maps a concept from the set $\mathcal{C}$ of all concepts to the set of neighbors of that concept which is a subset of the power set $\mathcal{P}(\mathcal{C})$ of all concepts in the ontology. The similarity of $\mathcal{C}_1$ and $\mathcal{C}_2$ is then computed as

$$\mathrm{DN}_{\mathcal{O}}(\mathcal{C}_1, \mathcal{C}_2) = \frac{1}{|\mathcal{C}_1| \cdot |\mathcal{C}_2|} \sum_{(c_1,c_2)^{\mathrm{T}} \in \mathcal{C}_1 \times \mathcal{C}_2} \frac{|n_{\mathcal{O}}(c_1) \cap n_{\mathcal{O}}(c_2)|}{|n_{\mathcal{O}}(c_1) \cup n_{\mathcal{O}}(c_2)|}. \tag{4}$$

The second similarity measure is based on the average number of hops between pairs of concepts and thus counts as an edge-counting measure. This very simple type of similarity measure was e.g. proposed in [17]. The assumption is that concepts are very similar if there exists a very short path between them in the ontology. The longer the shortest path between them the more different two concepts are assumed to be. To evaluate the paths between all remaining concepts in the sets $\mathcal{C}_1$ and $\mathcal{C}_2$ we first compute the sum

$$\mathrm{S}_{\mathcal{O}}(\mathcal{C}_1, \mathcal{C}_2) = \sum_{(c_1,c_2)^{\mathrm{T}} \in \mathcal{C}_1 \times \mathcal{C}_2} p_{\mathcal{O}}(c_1, c_2) \tag{5}$$

and the maximum

$$\mathrm{M}_{\mathcal{O}}(\mathcal{C}_1, \mathcal{C}_2) = \max_{(c_1,c_2)^{\mathrm{T}} \in \mathcal{C}_1 \times \mathcal{C}_2} \{p_{\mathcal{O}}(c_1, c_2) + 1\}. \tag{6}$$

Here, $p : \mathcal{C}^2 \rightarrow \mathbb{N}$ is a function that maps two concepts to the length of the path between them. From these values we compute the weighted average path length

$$\overline{\mathrm{len}_{\mathcal{O}}}(\mathcal{C}_1, \mathcal{C}_2) = \frac{N_1 \, \mathrm{S}_{\mathcal{O}}(\mathcal{C}_1, \mathcal{C}_2) + N_2 \, \mathrm{M}_{\mathcal{O}}(\mathcal{C}_1, \mathcal{C}_2)}{|\mathcal{C}_1| \cdot |\mathcal{C}_2|} \tag{7}$$

with the normalizing factors

$$N_1 = \sum_{(c_1,c_2)^{\mathrm{T}} \in \mathcal{C}_1 \times \mathcal{C}_2} 1_{\mathcal{O}}(c_1, c_2) \qquad N_2 = |\mathcal{C}_1| \cdot |\mathcal{C}_2| - N_1. \tag{8}$$

The indicator function $1_{\mathcal{O}}$ becomes 1 in case two concepts are connected by a path in the ontology and is 0 otherwise. With the average path length we define our *graph distance* similarity measure as

$$\mathrm{GD}_{\mathcal{O}}(\mathcal{C}_1, \mathcal{C}_2) = \exp\left(1 - \overline{\mathrm{len}_{\mathcal{O}}}(\mathcal{C}_1, \mathcal{C}_2)\right). \tag{9}$$

The exponential scaling ensures that the minimal average path length of 1 also yields a similarity of 1 and for longer paths the similarity decreases towards 0.

The third similarity measure is also based on paths in the ontology graph, however, instead of taking the direct route between pairs of concepts hierarchical substructures in the ontology graph are explicitly considered. This measure is also an edge-counting approach. To consider hierarchical substructures in the ontology graph the path length function $p_{\mathcal{O}}$ is substituted by a function $t_{\mathcal{O}} : \mathcal{C}^2 \to \mathbb{N}^2$ that yields the path lengths of a pair of concepts to their nearest common ancestor. Similar to Eqs. (5) and (6) we then compute the sum

$$\mathrm{S}_{\mathcal{O}}(\mathcal{C}_1, \mathcal{C}_2) = \sum_{(c_1,c_2)^{\mathrm{T}} \in \mathcal{C}_1 \times \mathcal{C}_2} (t_1 + t_2 - 2) \tag{10}$$

and the maximum

$$\mathrm{M}_{\mathcal{O}}(\mathcal{C}_1, \mathcal{C}_2) = \max_{(c_1,c_2)^{\mathrm{T}} \in \mathcal{C}_1 \times \mathcal{C}_2} \{t_1 + t_2 - 1\} \tag{11}$$

where $(t_1, t_2)^{\mathrm{T}} = t_{\mathcal{O}}(c_1, c_2)$. From these values we compute the average path length according to Eq. (7) and then get the *tree distance* similarity as

$$\mathrm{TD}(\mathcal{C}_1, \mathcal{C}_2) = \exp\left(-\overline{\mathrm{len}_{\mathcal{O}}}(\mathcal{C}_1, \mathcal{C}_2)\right). \tag{12}$$

The exponential scaling ensures that the minimal average path length of 0 yields a similarity of 1 and similarity values slowly decrease towards 0 for longer paths.

The last similarity measure is based on the *normalized dissimilarity* from [16]. The motivation behind this measure is that similar concepts are subsumed under the same parent concepts in an ontology. Thus, to assess the dissimilarity of two concepts the sets of their parent concepts are compared. Let $\phi : \mathcal{C} \to \mathcal{P}(\mathcal{C})$ be a function that yields the set of all parent concepts for each concept in the ontology. With the help of this function the normalized dissimilarity in $[0, 1]$ of two concepts $c_1, c_2 \in \mathcal{C}$ can be computed as

$$\mathrm{dis}_{\mathrm{norm}}(c_1, c_2) = \log_2\left(1 + \frac{|\phi(c_1) \backslash \phi(c_2)| + |\phi(c_2) \backslash \phi(c_1)|}{|\phi(c_1) \backslash \phi(c_2)| + |\phi(c_2) \backslash \phi(c_1)| + |\phi(c_1) \cap \phi(c_2)|}\right). \tag{13}$$

To transform this dissimilarity into a similarity we use $1 - \mathrm{dis}_{\mathrm{norm}}$ and to compare two sets of concepts we use the average of the individual values. Thus, we get

$$\mathrm{ND}(\mathcal{C}_1, \mathcal{C}_2) = \frac{1}{|\mathcal{C}_1| \cdot |\mathcal{C}_2|} \sum_{(c_1,c_2)^{\mathrm{T}} \in \mathcal{C}_1 \times \mathcal{C}_2} (1 - \mathrm{dis}_{\mathrm{norm}}(c_1, c_2)). \tag{14}$$

Any of the similarity measures defined in Eqs. (4), (9), (12), and (14) can now be substituted for $s_{\mathcal{O}}$ in Eq. (3). More details regarding the first three similarity measures can be found in [32].

After we have now described the ontologies and similarity measures used in this article the next section is concerned with the survey we conducted to evaluate the computed similarity values.

5 Survey

In this section we describe the survey we conducted to gather similarity values from humans.

As mentioned in Sect. 3 our application is concerned with 236 837 keywords which consist of 26 324 unique terms. Since these are too many terms to ask for human similarity assessments for all of them we selected a subset of 74 terms. Doing so we especially considered that many of the 26 324 terms are very domain specific (e.g. small fashion brands) and probably not known to the general public. Thus, we selected terms of which we thought that they are more widely known.

The 74 selected terms can be paired up to 2 701 term pairs. However, we have to consider that for each term there are only a few similar terms and very many completely unrelated terms. When conducting a survey it is tiresome and frustrating for the participants if most of the term pairs they are asked about consist of unrelated terms and they must select the answer "very dissimilar" over and over again. Thus, when selecting term pairs for the survey, pairs of similar terms should be selected with proportionally higher probability than pairs of unrelated terms. To achieve this we ranked the term pairs according to their similarity computed with the DN measure. This should not bias the obtained results since the ranking only influences the probability with which term pairs are selected and does not change the similarity values later assigned by the participants of the survey. We then created 28 buckets of equal width of similarity values and randomly drew pairs from each bucket without replacement. This way terms with high similarity according to DN are selected with higher probability than pairs of completely unrelated terms. In several test subjects this over-sampling of term pairs with relatively high similarity led to considerably less frustration when participating in the survey compared to a uniquely distributed sampling of term pairs.

In total, we drew 135 term pairs which were randomly separated into 3 surveys with 45 questions each. At the beginning of each survey the same four examples of fashion term pairs with suggestions of similarity values were given to each participant to explain what we were looking for. After this introduction we asked some demographic questions to see if correlations to our measures differ between different demographic groups. The first two questions asked about the gender (male/female) and the age (below 18, 18 to 23, 24 to 30, 31 to 40, 41 to 60, 61 to 80, above 80) of each participant. With the next two questions we wanted to ascertain the fashion expertise of the participants. We asked how well they assess their fashion knowledge themselves (very good, good, not so good) and how much money they spend each month on fashion items including shoes (up to 25€ , 25 to 50€ , 50 to 100€ , 100 to 150€ , 150 to 200€ , more than 200€).

After this introduction and general part each participant was presented with 50 pairs of fashion pairs, one pair at a time. For each pair the similarity should be assessed on a scale from 1 to 7. In case a participant did not know some term or did not feel comfortable evaluating the similarity they were allowed to skip the question. 5 of the 50 similarity questions were control questions. One question contained the same term twice to check whether or not the maximal similarity value of 7 was assigned in this case. Two questions repeated previous term pairs to see whether or not consistent answers were provided and two further questions repeated previous term pairs but with switched term order.

6 Evaluation and Discussion

In this section we evaluate and discuss the results of our survey.

Altogether 183 people participated in our survey. Many of them were students or research assistants but there are also several friends and family members. For the evaluation of the results we first performed several preprocessing steps to eliminate duplicate and unreasonable answers. For the first survey 2 answers were duplicates possibly caused by people clicking the submit button multiple times. 3 further participants failed to assign the maximum similarity value of 7 to the control question that asked for the similarity of a term to itself. For the other control questions that repeated previous term pairs a threshold has to be defined because many people do not assign the exact same similarity value several questions later because they cannot remember the exact value the assigned previously. Thus, we decided to allow a deviation of 1 similarity unit which still eliminated another 19 answers for the first survey. This means that for the first survey 37 answers remain for the final evaluation steps.

For the second survey 3 duplicate entries had to be eliminated. In this case no one failed the control question with identical terms and 14 participants failed the other control questions. This leaves 44 answers of the second survey for final evaluation. In case of the third survey we have 1 duplicate and 1 failure of the identical terms question. 13 answers were excluded due to the control questions with repeated term pairs. Thus, 46 answers remain of the third survey which makes a total of 127 answers over all three surveys.

To compare the survey results with similarities computed with our similarity measures we use Pearson's and Spearman's correlation coefficients. While Pearson's correlation only considers a linear dependency between dimensions Spearman's correlation compares a ranking of points and thus also captures some non-linear correlations. Pearson's correlation coefficient is defined as

$$\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}. \tag{15}$$

Here, the pair $(x_i, y_i)^{\mathrm{T}}$ contains the averaged similarity values assigned by humans and as second entry the automatically computed similarity.

To compute Spearman's correlation coefficient both the automatically computed and manually assigned similarity values are ranked. Then, Spearman's correlation is defined as

$$1-\frac{6\sum_{i=1}^{n}(r_{x,i}-r_{y,i})^2}{n(n^2-1)} \tag{16}$$

where $r_{x,i}$ is the rank of the average of the manually assigned similarity of the i-th term pair and $r_{y,i}$ is the rank of the corresponding automatically computed similarity.

In addition to the correlation coefficients we compute the *Distinguishability* of of each similarity measure. We define the Distinguishability as the fraction of term pairs which are assigned a unique similarity value by a measure among the tested set of term pairs. A high value of Distinguishability close to 1 means that many term pairs can be distinguished by just looking at their similarity values. A low Distinguishability close to 0 means that many term pairs are assigned the same similarity value and cannot be distinguished by the respective measure. We argue that it is desirable to be able to distinguish as many term pairs as possible since a similarity measure should capture even fine grained semantic differences. Also, in our application where we want to aggregate data associated to terms we would like to gradually add data and not add a whole bulk of data from a set indistinguishable terms in one go.

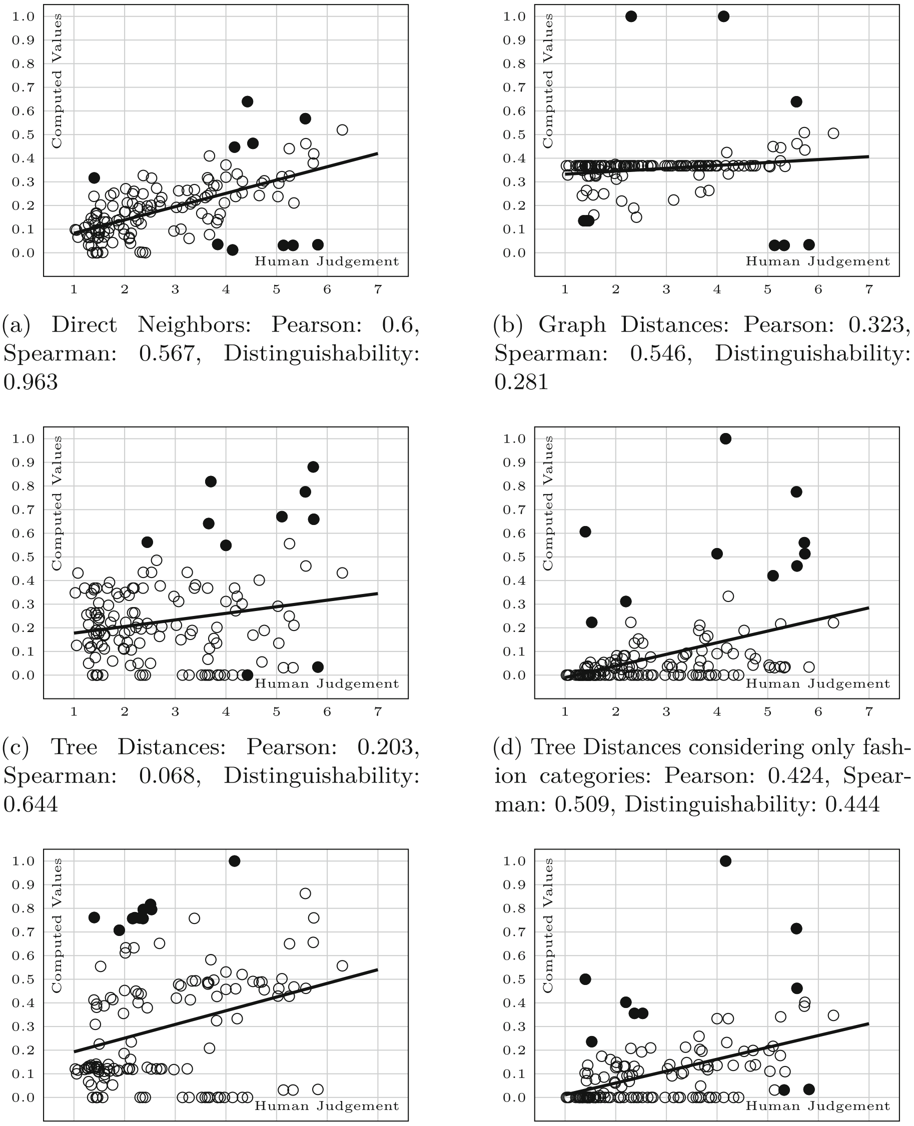
Of course for very large sets of terms the Distinguishability is bound to take values smaller than 1 because the probability for duplicate similarity values increases for very many term pairs. However, Distinguishability values of different similarity measures can still be interpreted relative to each other. Also, the set of 135 term pairs we used in our survey is small enough to allow unique similarity values for all of them.

6.1 Overall Correlations

In Fig. 1 correlation plots are given for all presented similarity measures and the averaged human judgement for the 135 term pairs. In addition to the similarity values the first principal axis is given by a straight line and the 10 similarity pairs with the greatest deviation from the axis are highlighted as solid black circles.

Figure 1(a) shows the results of the direct neighbors measure (DN). It achieves the highest correlation according to both a Pearson's correlation coefficient of 0.6 and a Spearman's correlation coefficient of 0.567. Although not the whole range of values occurs, especially there are no similarities above 0.65, the similarities are well scattered in the remaining range in contrast to some of the other measures where similarities agglomerate around certain values. Also, according to the Distinguishability this DN similarity measure performs best with a value of 0.963 which means that very few term pairs are assigned an identical similarity value.

The highlighted 10 term pairs whose similarities deviate the most from the shown principal axis are given in Table 2. They can be split in two groups, one with term pairs whose similarity is overestimated by the DN measure and one group whose similarity is underestimated by the measure. In cases where one of the terms occurs as a substring in the second term, the similarity gets underestimated. This is due to the fact that there exist many more nodes in the ontology which match to the shorter of the two terms and only some nodes which match to the more specific longer term. This fact leads to the neighborhoods of



(a) Direct Neighbors: Pearson: 0.6, Spearman: 0.567, Distinguishability: 0.963

(b) Graph Distances: Pearson: 0.323, Spearman: 0.546, Distinguishability: 0.281

(c) Tree Distances: Pearson: 0.203, Spearman: 0.068, Distinguishability: 0.644

(d) Tree Distances considering only fashion categories: Pearson: 0.424, Spearman: 0.509, Distinguishability: 0.444

(e) Normalized Dissimilarity: Pearson: 0.296, Spearman: 0.269, Distinguishability: 0.822

(f) Normalized Dissimilarity considering only fashion categories: Pearson: 0.449, Spearman: 0.515, Distinguishability: 0.474

Fig. 1. Correlation plots of similarity measures with crowdsourced human judgement. In addition to the similarity values the first principal axis is displayed and the 10 values with the greatest deviation from that axis are highlighted as solid circles.

Table 2. Term pairs whose similarity gets most over- resp. underestimated by the DN similarity measure in comparison with human judgement. Translations of the original German terms are given where appropriate for the reader's convenience. However, the original terms are left to assure the original semantics, which may have subtly changed due to the translation, can be reconstructed by the reader.

Overestimated	bergschuhe (mountain boots), winterstiefel (winter boots)
	abendkleid (evening dress), strickkleid (knit dress)
	kopfhörer (headphones), regenschirm (umbrella)
	mütze (cap), hüte (hats)
	klettverschluss (Velcro), reißverschluss (zipper)
Underestimated	calvin klein, nike
	sportbekleidung (sportswear), jack wolfskin
	schuhe (shoes), badeschuhe (flip flops)
	bergschuhe (mountain boots), schuhe (shoes)
	jacke (jacket), lederjacke (leather jacket)

the two terms not overlapping to a great degree and, thus, a low similarity assessment. However, overestimation does not happen in all cases of term pairs with such a characteristic. There are pairs, e.g. "hemd" and "unterhemd" (engl. "shirt" and "undershirt"), for which the similarity assessment of the DN measure more accurately mirrors the human judgement.

The two other cases in which the similarity is underestimated by the DN measure are concerned with names of fashion brands. In the first case two brand names ("calvin klein" and "nike") are compared. Here, humans assign a high similarity value based on the fact that both terms are brand names. However, it appears that the products associated to the two brands are rather different which is considered by the DN similarity measure.

Two cases in which the similarity of term pairs is overestimated by the DN measure are concerned with accessories. These are regarded as rather similar by the DN measure because the corresponding products are in similar categories and have similar properties. Humans, however, assign them with lower similarities because they are rather different products. The 10 terms which deviate most from the principal axis are not analyzed in detail for the other similarity measures for the sake of brevity. To summarize those results, we found that computed similarities and human judgement deviate most for similar types of terms.

Figure 1(b) shows the results of the graph distances measure (GD). Here, Pearson's correlation coefficient takes a distinctly lower value of 0.323 and also Spearman's correlation is a little bit lower with a value of 0.546. When considering the similarity values assigned by the GD measure we see that many pairs get a similarity of approx. 0.38. In fact the GD measure achieves a very low Distinguishability of 0.281. This is due to the fact that the corresponding nodes in the fashion ontology are connected by paths of similar lengths.

In Fig. 1(c) the results of the tree distances similarity measure (TD) are visualized. In comparison to the DN and GD similarity measures the results are much more scattered in this case. Also the TD measure achieves very low correlations of 0.203 (Pearson) and 0.068 (Spearman) which indicates very low or nearly no correlation to the human judgement. While there are some term pairs which are assigned identical similarity values, especially the value 0, the Distinguishability of 0.644 is rather high.

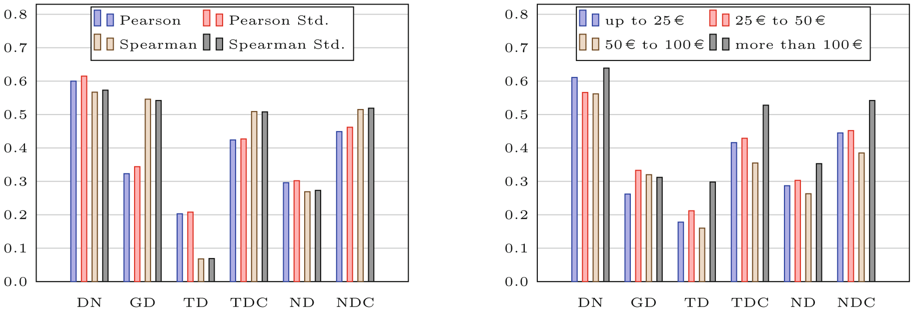
Since the TD similarity measure is based on tree structures it may be useful to restrict it to use only the fashion category hierarchy of the fashion ontology since this is it's main hierarchical structure (cf. Sect. 3). The resulting similarity measure is denoted by TDC and it's results are shown in Fig. 1(d). In comparison to the unrestricted TD measure many term pairs, especially those which are regarded as dissimilar by humans, are assigned distinctly lower similarity values. This leads to higher correlation with the human judgement expressed by a Pearson's correlation coefficient of 0.424 and a Spearman's correlation coefficient of 0.509 which are both higher than their respective counterparts of the unrestricted TD measure but not as high as those of the DN measure. However, restricting the tree distance measure to only the category hierarchy of the fashion ontology lowers the Distinguishability since now many term pairs are assigned with a similarity value of 0. Many of those pairs are concerned with fashion brands which obviously cannot accurately be compared using a hierarchy of fashion categories.

The next similarity measure whose results are visualized in Fig. 1(e) is based on the normalized dissimilarity (ND) measure from [26]. There, it was shown that this measure performs best among several others when applied to the WordNet ontology. The ND measure is similar to the TD measure since it is also based on the tree structure of the used ontology. The obtained results are also similar to those of the TD measure although both correlations and the Distinguishability are a little higher than with the TD measure. We get a Pearson's correlation coefficient of 0.296 and a Spearman's correlation coefficient of 0.269. Although these correlations are higher than those of the TD measure they are nowhere close to those of the DN measure. The Distinguishability of 0.822 is rather high although still smaller than that of the DN measure.

Similar to the TDC measure it is sensible to restrict the ND measure to only make use of nodes belonging to the fashion category hierarchy since they constitute the most hierarchical part of the fashion ontology. This measure is denoted by NDC in the remainder of this article. The results for the NDC measure are visualized in Fig. 1(f). Similar to the TD measure restricting the ND measure to fashion categories increases the correlation with human judgement but decreases the Distinguishability. We get a Pearson's correlation coefficient of 0.449 and a Spearman's correlation coefficient of 0.515. However, the Distinguishability drops from 0.822 to 0.474 due to many term pairs now getting a similarity of 0. These especially are term pairs with at least one brand name which obviously cannot accurately be compared using a hierarchy of fashion categories.

6.2 Correlations for Demographic Groups

In addition to the overall correlation we also evaluated the correlation between the similarity measures and several demographic groups of participants. The most interesting results are presented in Fig. 2.



(a) Standardized replies.

(b) Spending per month.

Fig. 2. Correlations of similarity measures and human judgement grouped by different properties of the participants of the survey. If not stated otherwise Pearson's correlation coefficient is used.

First Fig. 2(a) compares the overall correlations with those obtained by first standardizing the replies of each participant in such a way that their mean reply is 0 with a variance of 1. This makes results of the participants more comparable since different people make different use of the used rating scale and especially many people only make sparse or no use of the extreme value 1 and 7.

In most cases standardization leads to a very slight increase in the correlations with human judgement. Also, Fig. 2(a) visualizes well that in case of the GD measure Spearman's correlation coefficient is much higher than Pearson's coefficient independent of any standardization of replies. In case of the TD measure the opposite is the case. Here, Pearson's correlation coefficient is much higher than Spearman's coefficient. This suggests that in case of the GD measure the dependency between calculated similarities and human judgement has a non-linear component.

The most interesting of the demographic questions we asked (cf. Sect. 5) with regard to the correlation of human judgement and automatically computed similarity values is the amount of money participants spend per month on fashion items (including shoes). We asked this question because we proposed that it is reasonable for someone who spends much on fashion items to have a higher expertise in this area than someone who spends very little. Due to too few answers we had to aggregate all categories with more than 100€ spending volume. After the aggregation we have 26 participants who spend less than 25€ per month on fashion items, 38 participants who spend between 25€ and 50€, 46 participants who spend between 50€ and 100€, and 16 participants spend more than 100€ per month on fashion items. One participant did not answer this question.

Figure 2(b) shows the correlations grouped by monthly spending volume. The most noticeable difference in correlations is between the group spending more than 100€ per month and the other groups. Especially for the TD, TDC, ND, and NDC measures the correlation is distinctly higher for this group than for the other groups. Also, for these measures participants spending 50€ to 100€ per month achieve a slightly lower correlation than the other groups. For the DN measure the correlation is also slightly higher for the group spending more than 100€ whereas for the GD measure it is slightly lower than the correlation for the other groups. This result supports our proposition that, at least with respect to the used similarity measures, a high monthly spending volume indeed signifies a higher fashion expertise.

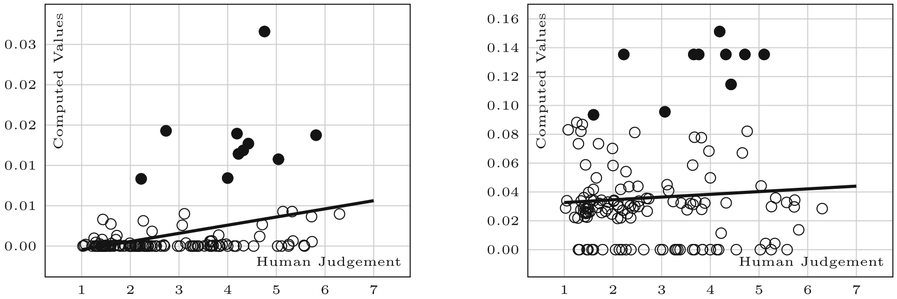
The results of the remaining demographic questions we asked are not presented in more detail for the sake of brevity. We found that among different age groups, genders, and self-assessed fashion expertise the correlation with the similarity measures is constant apart from small fluctuations.

6.3 DBpedia

To assess the influence of the ontology on the similarities we applied the similarity measures to a fragment of the DBpedia (cf. Sect. 3). Similar to Sect. 6.1 we computed the correlations between the measures applied to the DBpedia fragment and the averaged human judgement. However, in contrast to our rather small fashion ontology the fragment of the DBpedia even though being only a fragment is still too large to evaluate all of our measures. For the GD measure we had to limit the maximal depth for which paths are searched to 4. The tree based measures TD and ND take too long to evaluate even when limiting the search depth. Also, since the DBpedia fragment only contains one type of edge it makes no difference to restrict the tree based measures to only use edges of that type.

The results for the DN measure are visualized in Fig. 3(a) In contrast to the fashion ontology all computed similarity values are much smaller and several term pairs are assigned a similarity of 0. This is partially due to terms for which no matching node exists in the DBpedia fragment and partially due to nodes having no common neighbors in the DBpedia because nodes lie further apart in this large graph and only direct neighbors are considered. The correlation to human judgement is also a lot lower than was the case for the fashion ontology with a Pearson's correlation coefficient of 0.364 and a Spearman's coefficient of 0.304. The Distinguishability is also much lower than in case of the fashion ontology and reaches 0.526.

In Fig. 3(b) the results of the GD measure are shown. When applied to the DBpedia fragment this measure shows nearly no correlation with human judgement with a Pearson's correlation coefficient of 0.074 and a Spearman's coefficient of 0.02. Nevertheless, the Distinguishability of 0.711 is higher than that achieved with the fashion ontology and also than that of the DN measure in



(a) Direct Neighbors: Pearson: 0.364, Spearman: 0.304, Distinguishability: 0.526

(b) Graph Distances: Pearson: 0.074, Spearman: 0.02, Distinguishability: 0.711

Fig. 3. Correlation plots of DBpedia based similarity measures with crowdsourced human judgement. In addition to the similarity values the first principal axis is displayed and the 10 values with the greatest deviation from that axis are highlighted as solid circles.

conjunction with DBpedia. This shows that paths between several node pairs which share no common neighbors can be found in the DBpedia.

7 Conclusion and Outlook

In this article we presented a custom fashion ontology and several similarity measures. The similarity values were compared with the results of a survey we conducted to gather crowdsourced similarity values for a set of 135 term pairs. From preparing the survey we especially learned that it is important to include enough pairs of similar terms in order to not bore and frustrate participants.

We evaluated the results of the survey in several ways. First of all, we showed that our proposed direct neighbors (DN) similarity measure achieved the highest correlation with human judgement compared to several other state of the art similarity measures.

Furthermore, we analyzed which term pairs deviate most from the principal axis of measurements to find out in which cases the automatic similarities fit worst to the human judgement. From this we learned that all measures have problems in case one term is contained as a substring in the other term, if one term describes a fashion category and the other one a brand, and if terms are involved that describe accessories. In these cases human judgement and the knowledge captured by the fashion ontology deviate most.

In addition to the correlation analysis we calculated the Distinguishability to evaluate how many term pairs were assigned unique similarity values by a measure since we argue that it is important to be able to distinguish as many terms as possible with a similarity measure. With regard to the Distinguishability the DN similarity measure achieved the highest ranking since there are only very few term pairs which are assigned identical similarity values.

In our survey we asked several questions to divide participants into different demographic groups. The main result with regard to demographics is that the correlation between automatically calculated similarities and human judgement is higher than average for people that spend much money on fashion items. This can be expected and confirms that our measures reflect the similarity assessment of people who spend much money on fashion items and thus have a high expertise in the domain.

Finally, we applied two of the similarity measures to a fragment of the DBpedia to determine the influence of the ontology on similarity assessments. Due to the size of this ontology the tree based measures could not be evaluated and we could just compare the DN and GD approaches. Of those two, however, the DN measure achieved the highest correlation with human judgement.

In the future we would like to improve the mentioned shortcomings of the DN similarity measure and apply it do further ontologies like WordNet. We also want to actually use the similarity measures in our application to find training data for machine learning models in case no historical data is yet available.

References

1. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: the concept revisited. ACM Trans. Inf. Syst. **20**(1), 116–131 (2002)
2. Bruni, E., Tran, N.K., Baroni, M.: Multimodal distributional semantics. J. Artif. Intell. Res. **49**, 1–47 (2014)
3. WordNet. http://wordnet.princeton.edu/. Accessed 17 Dec 2015
4. Ghose, A., Yang, S.: An empirical analysis of search engine advertising: sponsored search in electronic markets. Manage. Sci. **55**(10), 1605–1622 (2009)
5. Kelly, B., Burka, K.: Enterprise paid media compaign management platforms 2015: a marketer's report. http://downloads.digitalmarketingdepot.com/MIR_1305_PPCamp2013_buyersguidelandingpage.html. Accessed 16 Dec 2015
6. Thumasathit, T.: Wag the dog: the tail of bid management. http://searchenginewatch.com/sew/opinion/2048496/wag-dog-the-tail-bid-management. Accessed 16 Dec 2015
7. Chatfield, C.: Time-Series Forecasting. Chapman and Hall/CRC, Boca Raton (2000)
8. Box, G.E.P., Jenkins, G.M., Reinsel, G.C.: Time Series Analysis: Forecasting and Control, 4th edn. Wiley, Oxford (2008)
9. Brockwell, P.J., Davis, R.A.: Introduction to Time Series and Forecasting, 2nd edn. Springer, New York (2002)
10. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Stat. Comput. **14**, 199–222 (2004)
11. Pai, P.F., Hong, W.C.: Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms. Electr. Power Syst. Res. **74**(3), 417–425 (2005)
12. Tay, F.E.H., Cao, L.: Application of support vector machines in financial time series forecasting. Omega **29**(4), 309–317 (2001)
13. Kalkowski, E., Sick, B.: Generative exponential smoothing models to forecast time-variant rates or probabilities. In: Proceedings of the International Work-Conference on Time Series (ITISE 2015), pp. 806–817 (2015)

14. Kalkowski, E., Sick, B.: Probabilistic generative models to; forecast time-variant rates or probabilities. In: Rojas, I., Pomares, H. (eds.) Time Series Analysis and Forecasting. Contributions to Statistics. Springer, New York (2015, to appear)
15. DBpedia. http://wiki.dbpedia.org. Accessed 16 Dec 2015
16. 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)
17. Rada, R., Mili, H., Bichnell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Trans. Syst. Man Cybern. **19**(1), 17–30 (1989)
18. Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense identification. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database, pp. 265–283. MIT Press, Cambridge (1998)
19. Li, Y., Bandar, Z.A., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans. Knowl. Data Eng. **15**(4), 871–882 (2003)
20. Al-Mubaid, H., Nguyen, H.A.: Measuring semantic similarity between biomedical concepts within multiple ontologies. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **39**(4), 389–398 (2009)
21. Tversky, A.: Features of similarity. Psycological Review **84**(4), 327–352 (1977)
22. Petrakis, E.G.M., Varelas, G., Hliaoutakis, A., Raftopoulou, P.: X-similarity: computing semantic similarity between concepts from different ontologies. J. Digital Inf. Manage. **4**(4), 233–237 (2006)
23. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 1995), vol. 2, Montreal, QC, Canada, pp. 448–453 (1995)
24. Cilibrasi, R.L., Vitányi, P.M.B.: The Google similarity distance. IEEE Trans. Knowl. Data Eng. **19**(3), 370–383 (2007)
25. Zhou, Z., Wang, Y., Gu, J.: A new model of information content for semantic similarity in WordNet. In: Second International Conference on Future Generation Communication and Networking Symposia (FGCNS 2008), vol. 3, Sanya, Hainan Island, China, pp. 85–89 (2008)
26. Sánchez, D., Montserrat, B.: Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective. J. Biomed. Inform. **44**(5), 749–759 (2011)
27. Milne, D., Witten, I.H.: An open-source toolkit for mining Wikipedia. Artif. Intell. **194**, 222–239 (2012)
28. Bollegala, D., Matsuo, Y., Ishizuka, M.: A relational model of semantic similarity between words using automatically extracted lexical pattern clusters from the web. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009), pp. 803–812 (2009)
29. Rubenstein, H., Goodenough, J.: Contextual correlates of synonymy. Commun. ACM **8**, 627–633 (1965)
30. Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. Lang. Cogn. Process. **6**, 1–28 (1991)
31. Neo Technology Inc: Neo4J. http://neo4j.com. Accessed 17 Dec 2015
32. Kalkowski, E., Sick, B.: Using ontology-based similarity measures to find training data for problems with sparse data. In: Proceedings of the 2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2015), pp. 1693–1699 (2015)