

How Similar Is It? Towards Personalized Similarity Measures in Ontologies

Abraham Bernstein, Esther Kaufmann, Christoph Bürki

Department of Informatics, University of Zurich, Switzerland

Mark Klein

MIT – CCS, Cambridge, MA, U.S.A.

Finding a good similarity assessment algorithm for the use in ontologies is central to the functioning of techniques such as retrieval, matchmaking, clustering, data-mining, ontology translations, automatic database schema matching, and simple object comparisons. This paper assembles a catalogue of ontology based similarity measures, which are experimentally compared with a “similarity gold standard” obtained by surveying 50 human subjects. Results show that human and algorithmic similarity predications varied substantially, but could be grouped into cohesive clusters. Addressing this variance we present a personalized similarity assessment procedure, which uses a machine learning component to predict a subject’s cluster membership, providing an excellent prediction of the gold standard. We conclude by hypothesizing ontology dependent similarity measures.

Keywords: Ontologies, Similarity Measures, Semantic Web, Human Assessment of Similarity

1 Introduction

Claudia is a geneticist who is about to write a paper. As a responsible scientist she wants to look up some information categorized in a large gene-ontology and formulates a query specifying a number of attributes/features of her gene and some of its relationships. She is especially excited about the prospect of finding some other, similar research that she will be able to cite in her paper. When executing the query, however, she is buried in hundreds of results...

This is a very typical situation. People querying databases oftentimes find themselves either *buried in results* to their queries or find *no results whatsoever*. A common approach to dealing with these problems is to rank the results of a query, in the case of too many answers, or return similar documents, when no precise matches to the query exist [BaRi99; BriPa98]. Both of these approaches require a measure of *similarity* between answers and queries. Finding a good measure of

similarity is, thus, crucial for providing a good retrieval performance. But not only retrieval of objects profits from good similarity measures. A variety of techniques, such as clustering, data-mining, semantic sense disambiguation, ontology translations, automatic database schema matching, and/or simple object comparisons rely on good similarity measures. Thus, similarity prediction algorithms are a central element in the semantic web, artificial intelligence, or computer science researcher’s toolbox.

The increased use of ontologies raises the question of an appropriate similarity measure for the use with ontologies or semantically enhanced applications. Most semantic-web systems, however, use traditional logic approaches where corresponding objects are determined by perfect matches and similarity (as opposed to equivalence or subsumption) isn’t used as a concept. Humans, on the other hand, typically have little difficulty in determining the intended meaning of ambiguous words, expressions, or even complex objects, whereas it is challenging to replicate this process computationally.

This paper investigates algorithms for determining the semantic similarity between objects in an ontology. In particular, it *experimentally compares a number of adapted or existing computational measures* (mainly taken from the computer linguistics/natural language processing domain) *with the human judgment of the similarity of instances in an ontology*.

As such, the contributions of this paper are the following: First, it assembles a catalogue of similarity measures used in ontologies and complements them with some additional, adapted measures from related domains such as natural language processing (NLP). Second, it compares those measures against a “similarity gold standard” within a large ontology (with over 5000 entries) established in an experiment with 50 subjects finding, among other things, that the quality of similarity assessment algorithms might be ontology dependent. Third, the experimental results provide a surprising insight into how the human understanding of similarity might require personalized and application-specific similarity measures. Last, it introduces a combined similarity measure that promises to capture the personalized preferences using a machine learning algorithm.

The paper is structured as follows: Next, we review the literature on object similarity and present the findings as a catalogue of ontology similarity functions. Then, we provide a detailed explanation of our experimental setup, present the results of the experiment, and discuss limitations of the presented study. We close with a discussion of related work and some ideas for future work.

2 Semantic Similarity in Ontologies

The question of similarity is a heavily researched subject in the computer science, artificial intelligence, psychology, and linguistics literature. In particular the information retrieval literature has a long tradition of looking at measures for the similarity between *documents* [BaRi99; SaMc83]. Those approaches typically take the single words (or word stems) of a document as features and operate on histogram vectors thereof usually ignoring the ontological relationships of the words.

There are essentially two ways to make use of the hierarchical ontology structure for determining the *semantic similarity between objects in an ontology*: the *edge based* approach and the *node based* approach. The traditional edge based approach estimates the distance/edge length between nodes [Lee⁺93; Rada⁺89]. The shorter the path from one node to the other, the more similar they are. The problem with this approach is that it relies on the notion that edges in a taxonomy represent uniform distances, i.e. it assumes that all semantic links are of equal weight. The newer node based approaches [Res99] typically use information content measures or information on object-part relationships to determine the conceptual similarity. The similarity between concepts is determined by the extent to which they share information.

In this section we will present five different distance measures, both node and edge based, that are derived from the literature and are adapted to the context of comparing complex objects in an ontology. As complex objects we define entities with attributes, attribute values, and relationships of which one might be a specialization (i.e., an *is-a* relationship denoting any type of subclassing). This definition, thus, subsumes both explicit ontologies such as WordNet [Mill⁺93], where the specialization relationship is explicitly defined, as well as ontologies, where this relationship is to be derived logically (e.g., using subsumption). Consequently, we consider complex objects such as classes and instances in a semantic web ontology or a programming language, entities and records in a relational or object-oriented database, as well as any other compound data structure. For each of the measures we will briefly explain its source and how its adaptation to complex objects works.

Ontology Distance

The most intuitive similarity measure of objects in an ontology is their distance within the ontology. Obviously, *sparrows* are more similar to *geese* than to *whales*. They also reside closer in the typical biological taxonomies. The calculation of the ontology distance is based on the specialization graph of objects in an ontology.

The graph representing a multiple inheritance framework is not a tree but a directed acyclic graph. In such a graph the ontology distance could be defined as the shortest path going through a common ancestor or as the general shortest path, potentially connecting two objects through common descendants/specializations. For the purposes of this study we decided to employ the former, common-ancestor based specification, which seems to better reflect the common sense understanding of the closeness of two objects in a taxonomy. The pseudo-code algorithm looks as follows:

1. gen_a = all transitive generalizations of the object A
2. gen_b = all transitive generalizations of the object B
3. from $gen_a \cap gen_b$ determine the most recent common ancestor (MRCA)
4. ontology distance = count the length of the path from A to MRCA to B

Information-theoretic Approaches

The problem of the ontology distance is that it is highly dependent on the construction of the ontology. The measure is, therefore, highly dependent on (oftentimes) subjective ontology engineering decisions. To address this problem researchers in the NLP domain have proposed measuring the similarity between two objects (in their case words) in an ontology (i.e., WordNet) in terms of information-theoretic entropy measures [Lin98; Res99].

Specifically, Resnik [Res95; Res99] argues that an object (i.e., word) is defined by the members of the class specified. When using an explicit ontology like WordNet the set of members is equivalent to the descendants (hyponyms) of an object (word). The information of a class is defined as the probability $P(.)$ of finding a use of the class or its descendents in a corpus (in the case of WordNet: the probability of appearance of a word or one of its hyponyms in a corpus). The entropy of a class is the negative log of that probability. Similarity is now defined as:

$$\text{sim}(A,B) = (2 * \log P(\text{MRCA}(A,B))) / (\log P(A) + \log P(B)), \quad (1)$$

where MRCA is the most recent common ancestor of classes A and B. Intuitively, this measure specifies similarity as the probabilistic degree of overlap of descendants between two objects. Modeling his evaluation on an experiment by Miller and Charles [MiCha91], which uses human subjects to rate the similarity between 30 noun pairs, Resnik shows that this information theory based method provides significant improvement (correlation 0.79) over traditional edge methods (correlation 0.60).

We can directly reuse this approach for complex objects resulting in the following algorithm:

1. U = the total number of objects (or uses)
2. Find the most recent common ancestor (MRCA) of A and B
3. $P(A) = (\text{number of uses of } A) / U$
4. $P(B) = (\text{number of uses of } B) / U$
5. $P(\text{MRCA}) = (\text{number of specializations of MRCA}) / U$
6. $\text{sim}(A,B) = (2 * \log P(\text{MRCA}(A,B))) / (\log P(A) + \log P(B))$

Note that most ontologies today don't come with large annotated corpora, but instead contain instances. We can, hence, compute the $P(A)$ as the number of instances of A divided by the total number of instances.

Vector Space Approaches

Vector space models are very common in information retrieval [BaRi99; SaMc83] or machine learning [Mitch97]. They represent each object as a vector of features in a k -dimensional space and compute the similarity by measures such as cosine or Euclidean distance. We adapted the vector space model to the complex object setting by representing it as a k -dimensional vector. Here k is the number of unique object attributes/relations with a given value of the object and the length of the k -th component of the vector is associated with the object part frequency in the objects. The similarity between two objects' vectors is now simply defined as their inner product. The pseudo-code algorithm is:

1. Determine vector x from the object parts of A
2. Determine vector y from the object parts of B
3. $\text{sim}(A,B) = |xy| / |x| * |y|$

As an example consider the object *chair*, which has four *legs* and one *back*, to which it has a *has-part* relation, as well as a room *office*, to which it has a *is-in* relation. The chair vector $[4, 1, 1]$ would represent the chair in the space with the dimensions $[\text{has-part_legs}, \text{has-part_back}, \text{is-in_office}]$. Obviously, this type of "vectorization" is problematic as it, for example, does not capture that the dimensions *has-part_legs* and *has-part_back* are (semantically) closer related to each other than to *is-in_office*. It does, however, have the advantage of being computationally cheap. We, therefore, decided to use this measure as one option out of a whole set of possible vectorizations. An exhaustive study of complex object similarity measures would have to consider other vector space encodings as they are currently discussed in the propositionalization of relational machine learning problems [DzeLa01] and is beyond the scope of this paper.

Edit Distance (Levenshtein Distance)

The similarity between strings is often described as the edit distance (also called the Levenshtein Distance [Lev66]), the number of changes necessary to turn one string into another. Here a change is typically defined as either the insertion of a symbol, the removal of a symbol, or the replacement of one symbol with another. In our case we do not need to compare strings but objects. Therefore, we calculate the number of transformation steps needed to turn one object into another object. In other words, we count the number of insert, remove, and replacement operations of attributes, attribute values, relationships, or relationship types. Note that we ignored the names of attributes and relationships. Names could be added but would complicate the distance computation.

In a first version we assume equal costs (=1) for each of the transformations. In an alternative implementation we weigh each transformation type with a value that represents the “real” costs. For example, is the replacement transformation comparable with a deleting procedure followed by an insertion procedure? Hence, we could argue that the cost function c would have the following behavior:

$$c(\text{deleting}) + c(\text{inserting}) \geq c(\text{replacing}) \quad (2)$$

Using this assumption we calculate the worst case for the cost of a transformation from A to B by replacing all object parts of A with object parts of B, then deleting the rest of object parts of A, and inserting additional object parts of B. The worst case cost is then used to normalize the edit distance. The overall algorithm looks as follows:

1. Determine parts (attributes/relationships) of A
2. Determine parts of B
3. Compute *number of transformation steps* (replace, insert, delete) from A to B
4. Compute *worst case cost* for the procedure
5. Relative edit distance = (number of transformation steps) / (worst case costs)

Full-text Retrieval Method (tfidf)

The probably most often used similarity measure comes from the information retrieval literature and compares two documents by using a weighted histogram of the words they contain [BaRi99; SaMc83]. Specifically, the “term frequency and inverse document frequency” weighing scheme (short tfidf) works as follows: it counts the frequency of occurrence of a term in a document in relation to the word’s occurrence frequency in a whole corpus of documents. The resulting word counts are then used to compose a weighted term vector describing the document. The similarity between the two documents is then computed as the cosine between their respective weighted term vectors. In our case we created a (text) document

for each object in the ontology. Every document contained the object name, its attributes, and a brief description of its relationships (similar to the descriptions shown in Figure 1). We then took the cosine between the tfidf-weighted word vectors generated from each of the object-describing documents computed by an off-the-shelf algorithm [McCal96] as the similarity.

3 Experimental Evaluation

The similarity measures introduced above provide a first catalogue of candidates for an ontology based similarity metric. All of them have been used in some form or another for related problems and, therefore, have the potential of being useful in the semantic-web domain. In order to assess their usefulness, however, we need to evaluate them against a “gold standard” of object similarity.¹ To that end we designed a detailed experiment in which human subjects were asked to assess the similarity between two objects. As Budanitsky and Hirst [BuHi01] found in a study comparing WordNet similarity measures human judgments give the best assessments of the “goodness” of a measure, a finding supported by Blok et al. [Blok⁺02]. This section will describe the experimental setup and the statistical evaluation of the results setting the stage for a discussion of the results in the next section.

Study Design

To establish our gold standard we first needed a suitable experimental setup. We found that the experiment described in Miller and Charles [MiCha91], which relies on human judgments, has become the benchmark in determining the similarity of words in NLP research (see [BuHi01; JarSz01; JiCon97; Lin98; Res95; Res99]). We reused their overall experimental design and adapted it to be usable for complex objects in an ontology as follows: First, we had to find a number of suitable object pairs from a large ontology. Then, we had to define an appropriate order in which those pairs were going to be presented to the subjects, who assessed the similarity of the pairs on a scale between one (totally dissimilar) and five (identical). After carefully testing the overall survey with some test subjects and complementing it with demographic questions, we called on three groups of subjects to fill out the survey. Last but not least, we carefully evaluated the answers statistically. We will now visit each of these steps in detail.

¹ Alternatively, we could have evaluated the measures in a realistic application for similarity measures, which would go beyond this paper’s scope. The analysis can be found in [Bern⁺04].

As the *underlying ontology* we chose the *MIT Process Handbook ontology* [Mal⁺03; Mal⁺99], which contains over 5000 organizational processes and has been carefully developed for over 10 years. The ontology has a number of advantages. Each process in the ontology has a variety of relationships to attributes, subprocesses, exceptions, etc. and also provides a detailed textual description, providing the subjects with multiple types of information about the processes. Furthermore, the ontology has been used in other semantic-web projects [GroPo02; KleBe04] and treats a domain of interest to researchers in the semantic-web field (services and their description). Finally, note that the subjects' ability to relate to the ontology content is crucial for the success of the experiment. Lord et. al [Lord⁺03], for example, had to forgo an evaluation with human subjects as experts in their application domain (biology) are difficult to find. Consequently, the Process Handbook was especially suitable, as it treats a domain (business processes/services) that most people can relate to. Unfortunately, the Process Handbook is sometimes confusing in that it, like WordNet, doesn't distinguish between instances and classes.

From the Process Handbook we selected *40 processes* that we thought would be understandable to a general audience and *combined them into pairs* fulfilling the following criteria:

- At least one pair should be in close vicinity in the ontology-graph.
- At least one pair should be far apart in the ontology-graph.
- At least one pair should consist of a process and its descendant/specialization.
- One process was paired with itself.

The rest of the processes were paired in a way such that the processes' name, description, attributes, or relations (e.g., parts) featured some similarities.

Each pair was then turned into a web-page using the on-line survey tool OpinioTM, which offered a comfortable graphical user interface and permitted an accurate definition of survey parameters. As can be seen in Figure 1 the subjects were asked to assess the *similarity between two processes on a scale from 1 (no similarity) to 5 (identical)*. With a simple drop-down list the users could specify how they had made the assessment: 1. by process name, 2. by process description, 3. by process parts/relationships, 4. a combination of 1-3, and 5. using other assessment method. This question should capture in respect to which features of the object the similarity was observed by the subjects – a notion that similarity researchers in the social sciences have found to be central [GeMe98]. Finally, the subjects could add some comments on their assessment.

When participating, a subject was presented with a carefully arranged step-by-step introduction and was given the opportunity to assess a simple example. At the end of her assessment, she was offered to finish the survey or continue assessing a second group of ten pairs. When finishing the survey, the subjects were presented


with a final page of questions asking some *demographic questions* such as age group (e.g., 10-19, 20-29, ...), education (high-school, bachelor, ...), knowledge of English (none, basic, good, ...), and whether they had any knowledge in computer science (yes/no) or linguistics (yes/no). As usual we piloted it with test candidates.

| Survey | |
|---|---|
| Do you need help with the similarity assessment? Look at the example or check the FAQ-page . | |
| Process A Name: Acquire Description: The acquisition of resources that will be used to produce a product or service. At the most generic level, this process decomposes into only the identification of a need, when to get it and the getting of the input that meets the need. Communication of this need from the process that requires an input to the process that provides the output is not required, though it is present in many specializations. | Process B Name: Buy in a store Description: Buying in a store involves the buyer going to a physical or virtual location and purchasing an item/service. |
| Process parts: <ol style="list-style-type: none"> 1. Identify needs or requirements 2. Receive physical resource 3. Determine timing | Process parts: <ol style="list-style-type: none"> 1. Identify potential sources 2. Manage suppliers 3. Receive 4. Place order 5. Pay 6. Select supplier 7. Identify own needs |

The assessment is based on: ▼

No similarity ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 identical

Comments (optional):

 10%

Next

Figure 1: A sample survey page showing the two processes to compare

We deliberately *recruited our subjects from three different groups*. The first group consisted of the staff, researchers and faculty of the *computer science* department. The second group was chosen from the computer science *students* association. And the last group was made up from *computer linguists* (both students and staff). From each group an approximately equal number of subjects participated totaling in 50 survey participants, larger than any other study we found in the literature.

First Results/Data Analysis

To assess the quality of the similarity algorithms we compared their assessments with those of the subjects. This turned out to be a non-trivial task. First, while the algorithms provided nominal predictions the subjects' assessments were on an

ordinal scale. Second, the prediction of some algorithms was non-linear complicating their comparison using traditional correlations. We, therefore, chose to *compare each pair of assessments (including both the subjects' assessments and the algorithms' predictions) using the corrected Spearman correlation coefficient r_s* , which compares bindings (corrected ranks) of assessments rather than absolute values addressing both issues [Sachs02]. This coefficient compares two paired sets by assigning each number a rank with respect to its set and provides a number r_s between -1 and 1, where 1 represents perfectly correlated sets, -1 inversely correlated sets, and 0 completely uncorrelated sets. Typically values of $r_s \geq 0.5$, respectively $r_s \leq -0.5$, are taken as some correlation and values of $r_s \geq 0.7$, respectively $r_s \leq -0.7$, as good correlations. In other words, we took each series of similarity assessment (by either of the 50 human subjects or 5 algorithms) and compared it to every other assessment using the corrected Spearman rank correlation. The resulting correlation coefficients are represented in Figure 2 as grey-scales, in which the subjects are numbered from 2 to 51 and the algorithms have alphabetic identifiers (A, ..., F).

At a first glance the result looks rather abysmal. A large part of the assessments don't seem to correlate at the $r_s \geq 0.5$ (respectively ≤ -0.5) level. After careful consideration, however, we find the following interesting results in the data.

First, in general the algorithms seem to correlate no better or worse with the subjects' predictions than the subjects do among themselves. It even turns out that *the correlations of the subjects among each other are significantly similar to the correlations of each algorithm with the human subjects' assessments* (as shown by a t-test at level below 0.005 for all but one algorithm; below 0.9 for the information theory measure). Consequently, given the problematic correlation between the subjects' answers, the algorithms mostly perform significantly similar to "yet another subject," potentially (as we will see below) providing a good basis to mimic the human similarity measure.

Second, the weighted edit distance (A), simple edit distance (B), and vector space (C) predictions seem to correlate well with each other as well as (slightly less consistently) the information theory (D), ontology distance (E), and full-text (F) algorithms. After further consideration, we find that those two groups are indicating clusters in the answers of the subjects. The first cluster, identified by good correlation with either edit distance measures, is shown in Figure 3a. With few exceptions those subjects correlate above the $r_s = 0.5$ threshold. The second cluster, shown in Figure 3b, shows a similar inner cohesion. Consequently, *the subjects can be divided into clusters each showing a very high correlation with some of the measures*.

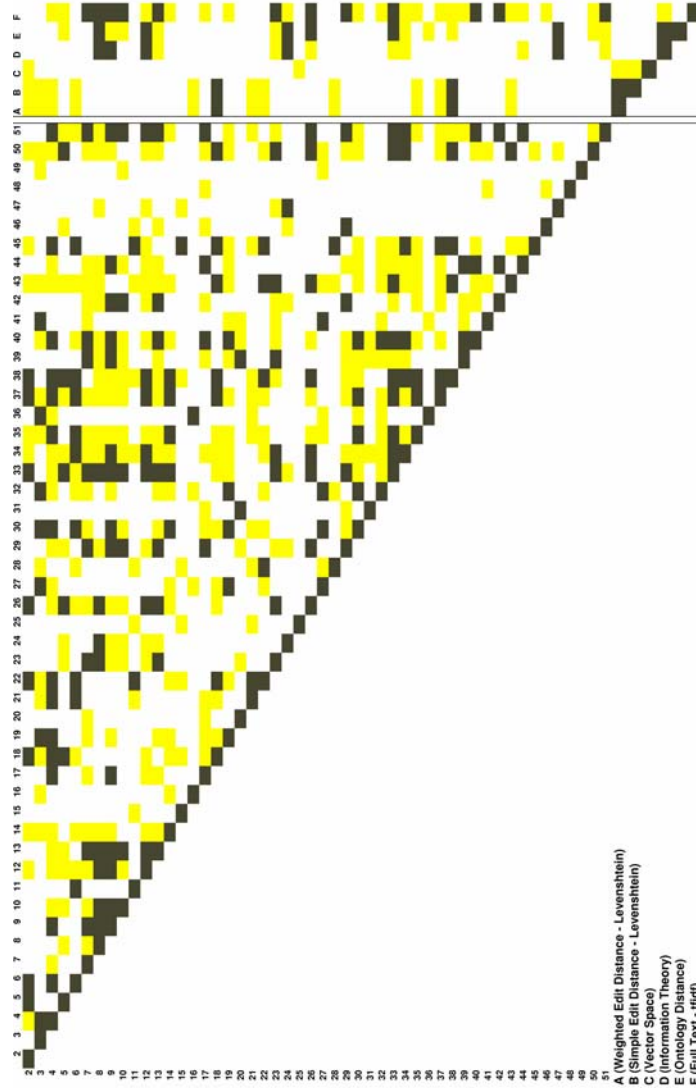


Figure 2: Grey-scale Spearman correlation matrix between all subjects (with numerical identifiers) and (right of the boxed line) the algorithms (with alphabetic identifiers). The fields are colored *white* for $\text{absolute}(r_s) < 0.5$, *grey/yellow* for $0.5 \leq \text{absolute}(r_s) < 0.7$, and *black* for $0.7 \leq \text{absolute}(r_s)$.

This is an important empirical finding: it shows that a general similarity measure reflecting human similarity assessments can hardly be found. Much more *widely applicable similarity measures will have to be personalized to the user's similarity assessment style*. While one might argue that those personalized measures are not

necessary for optimally completing purely computational tasks, they are likely to be more suitable when users are involved. This finding also provides rationale for the recent surge of personalized web search services by companies such as Google™ and Eurekster™.

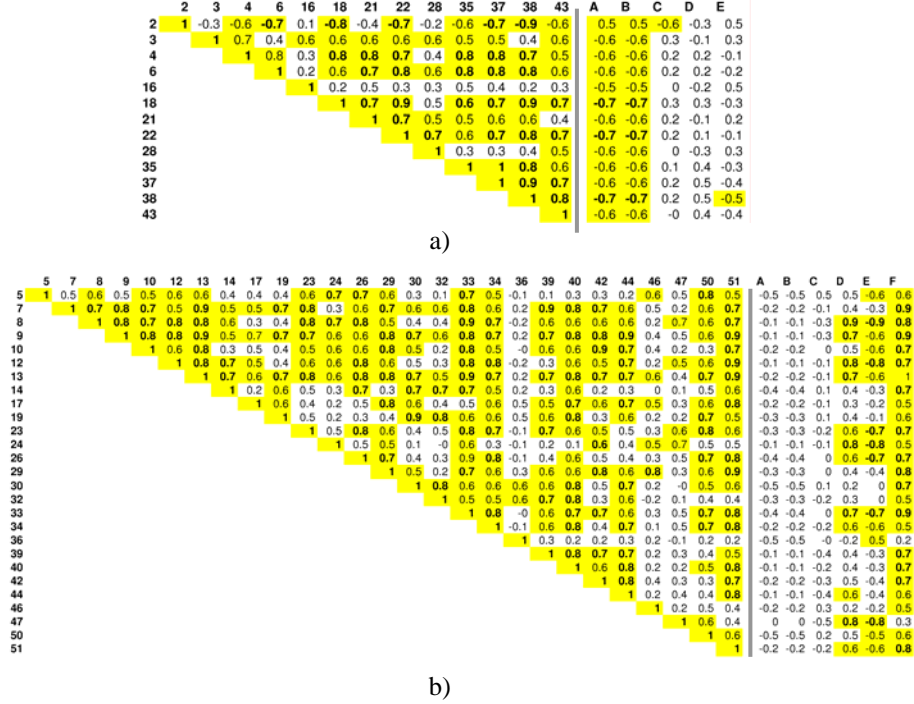


Figure 3: Corrected Spearman r_s for a) edit distance based cluster, b) ontology-oriented cluster. Values ≥ 0.5 (or ≤ -0.5) are highlighted yellow/grey; values ≥ 0.7 (or ≤ -0.7) are shown in bold.

Last, we find some *interesting results regarding the overall predictive quality of the measures*. The two edit distances performed equally, alleviating the need for any type of weighting. The vector space method, so successful in the full-text retrieval domain, performed rather disappointingly throughout. The information-theoretic method didn't perform too well. It was mostly equivalent in performance with the ontology distance method. This contradicts the findings from the NLP domain where information theoretic measures outperform the ontology distance [BuHi01; JarSz01; JiCon97; Lin98; Res95; Res99]. One explanation is the difference of the ontology underlying the experiment: the NLP findings based their experiments on WordNet; we used the Process Handbook. *This indicates that the quality of a similarity measure seems to be dependent on the ontology* – a question definitely deserving further attention in future research. The most

consistently performing method seems to be the adapted tfidf-measure. We will have to revisit the overall good performance of this measure below.

Summary

In this section we presented our adaptation of the well established semantic similarity experiment by Miller and Charles [MiCha91]. We asked 50 subjects from three different populations to assess the similarity between carefully chosen process pairs from within the large Process Handbook ontology. When comparing the assessed similarities using the corrected Spearman's rank correlation we found that (1) the algorithms correlated with the subjects' assessments to the same extend as the subjects did among themselves, that (2) the algorithms and human assessments could be grouped into cohesive clusters, and (3) we raised the question whether the applicability of a similarity measure is highly ontology dependent, as some measures unexpectedly outperformed others contradicting findings with WordNet.

4 Discussion: Towards a Personalized Similarity Predictor

The survey results provide an interesting foundation for further exploration: we have seen that the algorithmic similarity measures indeed mimic human similarity assessments as long as they belong to the same cluster. While this is a great success in itself it would be interesting to understand the nature of the two major clusters (shown in Figure 3). In this section we will first analyze the nature of the two clusters. Basing on this analysis we will then try to predict cluster membership, which will build the basis for a combined, personalized similarity algorithm.

What is the nature of the clusters?

The first question we will need to address is whether there seems to be a theoretical justification for the two arising clusters. Indeed, *the first group of measures (the two edit distances and the vector model) largely focuses on the object's parts*, i.e., its attributes and relationships (especially the subprocesses). The correlation between the two edit distances is to be expected; essentially they only differ in the weights. The vector space model is highly similar in that it builds a vector from the parts and uses the similarity between those vectors to assess similarity. Two measures of *the second group of measures, the information-theoretic approach and the ontology distance are oriented towards a process's location in the ontology*: the information-theoretic approach through its reliance

on the processes’ descendants, which are likely to be more common for closer objects, and the ontology distance by the direct count of the closeness in the ontology. Also recall that these two measures did correlate in the NLP experiments. Assuming that the Process Handbook ontology is well constructed and, therefore, reflects the “general” mental ontological model of the object/process pairs we chose this would lead to the following conclusions: Subjects who mainly assessed the similarity by the processes’ composition correlate with the first group. Subjects who mainly assessed the similarity using a (potentially implicit) ontology that possibly somewhat reflected the Process Handbook largely correlate with the second group.

But where does this explanation leave the full-text tfidf based method? As we can see in Figure 3, this measure mostly correlated with the second group and was the most consistently performing method. The nature of its good performance is likely to be found in its construction. Recall that this method based on a full-text description of the name of the object, all its attributes and their values, as well as all its relationships including the name of the objects participating in that relationship. In addition, it is useful to note that all the attributes (including the object’s description) and parts are inherited down the ontology. Therefore, they are likely to be similar for an object and its descendant unless they were changed by the ontology designers (who often rely on the ontology’s inheritance feature). As such this measure somehow combined the parts-oriented nature of the first cluster with the ontology-oriented nature of the second one.

What about the subjects that don’t seem to correlate with either cluster? We hypothesize that there are three explanations for those. First, they could be people who often changed their assessment method within the experiment. Alternatively, they could be people who mainly based their assessment on the processes’ names – an approach for which we didn’t have a similarity measure. Finally, there were some subjects who obviously only gave serious answers to the first few questions and then clicked through to the end of the survey.

Can we predict cluster membership?

The average (absolute) Spearman correlations of the algorithms, between 0.026 (for the vector measure) and 0.485 (for the tfidf method), are low, largely because of the clustering of the subjects. If we could predict a subject’s cluster membership, then we could just choose the best performing algorithm in that cluster and use that as the automated similarity measure. This would result in a combined, personalized similarity measure.

To assess whether this approach would actually work we first manually determined each subject’s cluster-membership. *For each subject we then chose the algorithm that performed best within the predicted cluster.* The result was an average correlation of 0.689, a highly significant result at the 5% level.

This motivated us to try to automatically predict a subject's cluster-membership. To that end we used the off-the-shelf decision tree learner J48, the Weka [WiFra00] implementation of C4.5 machine learning algorithm [Quin93]. As an input to the algorithm we used the (rather limited) demographic information gathered at the end of the survey. We complemented this information with the subjects' self-reported explanations of how they performed the similarity assessments. We then evaluated the quality of the algorithm's cluster-membership-prediction using a leave-one-out approach, which assesses how well a subject's cluster-membership could be predicted, given that the membership of all other subjects is known – a realistic setup for our problem. Even though our demographic data was so limited J48 could predict the cluster membership with an astonishing 70% accuracy. Furthermore, choosing the best performing algorithm in the J48-predicted cluster resulted in an average Spearman rank correlation of 0.624 – a highly significant figure (at the 5% level), only slightly worse than what we got with the manual cluster prediction. Confirming our hypothesizing about the nature of the clusters above, the algorithm found that the most discriminating feature for predicting a subject's cluster membership was whether (s)he had reported more than three times that (s)he had used the processes' parts as the major guiding principle when assessing similarity.

Summarizing, we found that the clusters indeed seem to be the result of different human similarity judgment processes. This indicates, again, that *we need to know more about human understanding of complex objects and ontologies in order to be able to devise appropriate algorithms for human-computer interaction*. Furthermore, we showed that *the use of a simple machine learning algorithm can provide the means* for deciding which of the presented algorithms to use, which could be used *to build a highly accurate and personalizable similarity assessment algorithm* – the goal we set ourselves at the onset of this paper.

The primary limitation of our evaluation is its restriction to one ontology. Assessing the generalizability of our findings requires the replication and augmentation of our experiment with other large ontologies. This is especially important as we found that methods that were found to be very predictive in other ontologies (such as WordNet) performed rather poor in our ontology. As mentioned above, however, finding a large ontology to which subjects can relate to is a difficult task, which we intend to undertake in a future study. Furthermore, we are convinced that additional work is needed to confirm the hypotheses regarding the nature of the clusters. Other ontologies could even give rise to additional clusters. This would definitely require researchers to gather more data about the subjects' assessment process. Last, a few subjects reported that they changed their assessment method during the test. How would an algorithm look like that could dynamically predict the similarity method desired for the next use of the retrieval, clustering, or other technique? What type of contextual input information would it require?

5 Related Work

The NLP literature provides the largest group of related work. Motivated by Resnik’s study [Res95; Res99] a number of papers describe improvements to his information-theoretic measure. Wu and Palmer [WuPal94] focus on the semantic representation of verbs in computer systems and find those measures well applicable in machine translation. Jiang and Conrath [JiCon97] propose a combined edge counting and node based method that outperforms either of the pure approaches. This hints at the usefulness of combined approaches like the cluster-aware one we proposed in the previous section.

Budanitsky and Hirst [BuHi01] support our claim that the quality of similarity measures is dependent on the ontology. They mention that differences in the quality of WordNet based assessment algorithms found in various papers can be explained by different versions of WordNet used. Jarmasz and Szpakowicz [JarSz01] empirically support this statement by showing how similarity measures based on the *Penguin’s Roget’s Thesaurus of English Word and Phrases Thesaurus* outperform those based on WordNet. Addressing this issue Lin [Lin98] tries to find an information-theoretic measure of similarity that is not tied to a particular domain or application and that is less heuristic in nature. The measure is found to outperform Resnik’s similarity algorithm slightly. It does, however, still require a probabilistic model of the application domain, which he gets from parsing a large word corpus. This limitation makes it problematic for smaller ontologies. Note that most of these approaches are focused on the comparison of *nouns*, limiting their generalizability to complex objects or even hierarchies of verbs (which the Process Handbook is in some sense).

Di Noia et al. [DiNoi*03] compare a human based ranking (20 subjects) of 12 items with the returns of an ontology based retrieval engine, which attains imprecise matching by relaxing query constraints. This is similar to using an ontologized edit distance for ranking retrieved objects. They find the automated rankings to show “...good correspondence...” to the average human subject’s assessment and refer to ongoing large-scale experiments for further details. Their work differs from ours in the focus on ranking retrieved objects rather than similarity measures in general. Furthermore, they do not compare their ranking method with any other approaches.

Using an experiment with 37 subjects Rodriguez and Egenhofer [RodEg03] find that feature matching is important for detecting the similarity of objects across ontologies relaxing the requirement for a single ontology. Their feature matching algorithm uses a weighted string matching operation of the words describing the feature, which is similar to a (specially) weighted string-oriented edit distance metric. Their study as well as the work of Wu and Palmer shows the potential that similarity measures have for supporting translations between ontologies.

Focusing on the bioinformatics application domain, Lord et al. [Lord⁺03] compare sequence similarity of proteins with Resnik's information-content based similarity operating on protein annotations. They found a good correlation between the two, but did not perform any subject based experiment due to the difficulty of obtaining domain-qualified subjects.

Ouzzani and Bouguettaya's [OuzBou04] propose and implement a generic approach for optimally querying web services using exact, overlapping, partial, as well as combined partial and overlapping matches on their input/output parameters. This is similar to a specially weighted edit distance matching over those parameters, whose sole use for retrieval has been shown to be problematic [KleBe04]. They don't report any evaluation of their approach. Andreasen et al. [And⁺03] discuss different principles for measuring similarity of atomic or compound concepts based on edge based principles extending the simple ontology distance metric we used. They don't report any evaluation or comparison to other similarity metrics.

Summarizing, we can say that we found no study that compared a comparable catalogue of similarity measures using a similar size subject pool as we did. While quite a few papers mention the need for ontology-specific measures, none of them seems to have found person-to-person differences. This could be due to the use of WordNet in most human subjects based experiments, which has been modeled after common sense use of the language opposed to most other ontologies, which are designed by specialists for a particular use.

6 Conclusions

In this paper we argued that similarity measures in ontologies, a central component of techniques such as clustering, data-mining, semantic sense disambiguation, ontology translations, automatic database schema matching, and simple object comparison, deserve more attention. We assembled a catalogue of five algorithms (one of which was presented in two versions) and compared them with an experimentally derived gold standard, which we obtained by surveying 50 human subjects. We found that human predictions had a large variance, but that the algorithms varied with them almost mimicking the subjects. We also found that the users and algorithms could be grouped into cohesive clusters showing that similarity assessments will have to be personalized to attain good results. We then constructed a personalized similarity assessment algorithm that predicts a subject's cluster membership using a machine learning algorithm providing surprisingly accurate similarity assessments for the subjects in our study. Last, given the difference of our results with the findings reported in the NLP literature, we hypothesized that the prediction quality of similarity assessment algorithms might be ontology dependent.

This study provides a first investigation of similarities in ontologies. Nevertheless, the task of understanding similarity in ontologies is far from over. To that end both technical work on better, feature combining, ontology-adapting, and personalized similarity assessment algorithms as well as behavioral studies exploring people's understanding of similarity and their use of similarity based features are needed.

7 Acknowledgements

The authors would like to thank the MIT Process Handbook project for making available the data on which the experimental evaluation is based, and the anonymous reviewers for their helpful comments. This work was partially supported by the Swiss National Science Foundation grant 200021-100149/1.

References

- [And⁺03] Andreasen, T.; H. Bulskov; Knappe, R.: From Ontology over Similarity to Query Evaluation. 2nd CoLogNET-ElsNET Symposium - Questions and Answers: Theoretical and Applied Perspectives. Amsterdam, Holland, 2003: pp. 39-50.
- [BaRi99] Baeza-Yates, R.; Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press; Addison-Wesley: New York; Harlow, England; Reading, Mass., 1999.
- [Bern⁺04] Bernstein, A.; Kaufmann, E.; Bürki, C.; Klein, M.: Object Similarity in Ontologies: A Foundation for Business Intelligence Systems and High-performance Retrieval. Twenty-Fifth International Conference on Information Systems. Washington, DC, 2004: pp. 11-25.
- [Blok⁺02] Blok, S.; Medin, D.; Osherson, D.: Probability from Similarity. AAAI Conference on Commonsense Reasoning. AAAI Press: Stanford, CA, 2002.
- [BriPa98] Brin, S.; Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Seventh International World Wide Web Conference. ACM-Press: Brisbane, Australia, 1998.
- [BuHi01] Budanitsky, A.; Hirst, G.: Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. Second meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001). Pittsburgh, PA, 2001.
- [DiNoi⁺03] Di Noia, T.; Di Sciascio, E.; Donini, F. M.; Mongiello, M.: A System for Principled Matchmaking in an Electronic Marketplace. WWW2003. Budapest, Hungary, 2003.

- [DzeLa01] Dzeroski, S.; Lavrac, N.: Relational Data Mining. Springer: Berlin; New York, 2001.
- [GeMe98] Gentner, D.; Medina, J.: Similarity and the Development of Rules. *Cognition* 65, 1998: pp. 263-297.
- [GroPo02] Grosz, B.; Poon, T. C.: Representing Agent Contracts with Exceptions using XML Rules, Ontologies, and Process Descriptions. International Workshop on Rule Markup Languages for Business Rules on the Semantic Web (held at ISWC2002). Sardinia, Italy, 2002.
- [JarSz01] Jarmasz, M.; Szpakowicz, S.: Roget's Thesaurus and Semantic Similarity. International Conference on Recent Advances in Natural Language Processing (RANLP2003). Borovets, Bulgaria, 2003.
- [JiCon97] Jiang, J. J.; Conrath, D. W.: Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. International Conference on Research on Computational Linguistics (ROCLING X). Taiwan, 1997.
- [KleBe04] Klein, M.; Bernstein, A.: Towards High-Precision Service Retrieval. *IEEE Internet Computing* 8, 2004: pp. 30-36.
- [Lee⁺93] Lee, J. H.; Kim, M. H.; Lee, Y. J.: Information Retrieval Based on Conceptual Distance in IS-A Hierarchies. *Journal of Documentation* 49, 1993: pp. 188-207.
- [Lev66] Levenshtein, V. I.: Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* 10, 1966: pp. 707-710.
- [Lin98] Lin, D.: An Information-Theoretic Definition of Similarity. Fifteenth International Conference on Machine Learning (ICML'98). Morgan-Kaufmann: Madison, WI, 1998.
- [Lord⁺03] Lord, P. W.; Stevens, R. D.; Brass, A.; Goble, C. A.: Investigating semantic similarity measures across the Gene Ontology: The relationship between sequence and annotation. *Bioinformatics* 19, 2003: pp. 1275-1283.
- [Mal⁺03] Malone, T. W., K. Crowston, and G. A. Herman (Eds.). 2003. Organizing Business Knowledge: The MIT Process Handbook. Cambridge, MA: MIT Press. 2003.
- [Mal⁺99] Malone, T. W.; Crowston, K.; Lee, J.; Pentland, B.; Dellarocas, C. et al.: Tools for inventing organizations: Toward a handbook of organizational processes. *Management Science* 45, 1999: pp. 425-443.
- [McCal96] McCallum, A. K.: Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. Unpublished manuscript (available online at: <http://www.cs.cmu.edu/~mccallum/bow/>), 1996, Download: 2004-04-15.
- [Mill⁺93] Miller, G. A.; Beckwith, R.; Fellbaum, C.; Gross, D.; Miller, K.: Introduction to WordNet: An On-line Lexical Database. Technical Report. Cognitive Science Laboratory, Princeton University, Princeton, NJ, 1993.
- [MiCha91] Miller, G. A.; Charles, W. G.: Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes* 6, 1991: pp. 1-28.
- [Mitch97] Mitchell, T. M.: Machine Learning. McGraw-Hill: New York, 1997.

- [OuzBou04] Ouzzani, M.; Bouguettaya, A.: Efficient Access to Web Services. *IEEE Xplore: Internet Computing* 8, 2004: pp. 34-44.
- [Quin93] Quinlan, J. R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann: San Mateo, CA, 1993.
- [Rada⁺89] Rada, R.; Mili, H.; Bicknell, E.; Bletner, M.: Development and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man, and Cybernetics* 19, 1989: pp. 17-30.
- [Res95] Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *14th International Joint Conference on Artificial Intelligence*. Montreal 1995: pp. 448-453.
- [Res99] Resnik, P.: Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research* 11, 1999: pp. 95-130.
- [RodEg03] Rodriguez, M. A.; Egenhofer, M. J.: Determining Semantic Similarity Among Entity Classes from Different Ontologies. *IEEE Transactions on Knowledge and Data Engineering* 15, 2003: pp. 442-456.
- [Sachs02] Sachs, L.: *Angewandte Statistik*. Springer: Berlin, 2002.
- [SaMc83] Salton, G.; McGill, M. J.: *Introduction to modern information retrieval*. McGraw-Hill: New York, 1983.
- [WiFra00] Witten, I. H.; Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan-Kaufmann: San Francisco, 2000.
- [WuPal94] Wu, Z.; Palmer, M.: Verb Semantics and Lexical Selection. *32nd Annual Meeting of the Associations for Computational Linguistics*. Las Cruces, New Mexico, 1994: pp. 133-138.