SimPack: A Generic Java Library for Similarity Measures in Ontologies

Abraham Bernstein, Esther Kaufmann, Christoph Kiefer and Christoph Bürki
Department of Informatics
University of Zurich
{bernstein,kaufmann,kiefer}@ifi.unizh.ch

July 18, 2006

Abstract Good similarity measures are central for techniques such as retrieval, matchmaking, clustering, data-mining, ontology translations, automatic database schema matching, and simple object comparisons. Measures for the use with complex (or aggregated) objects in ontologies are, however, rare, even though they are central for semantic web applications. This paper first introduces SimPack, a library of similarity measures for the use in ontologies (of complex objects). The measures of the library are then experimentally compared with a similarity "gold standard" established by surveying 94 human subjects in two ontologies. Results show that human and algorithm assessments vary (both between people and across ontologies), but can be grouped into cohesive clusters, each of which is well modeled by one of the measures in the library. Furthermore, we show two increasingly accurate methods to predict the cluster membership of the subjects providing the foundation for the construction of personalized similarity measures.

Paper Type: Working Paper

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 semantically annotated gene-ontology. She is especially excited about the prospect of finding some related work. When executing the query, however, she is buried in hundreds of results.

This is a very typical situation. People querying the web 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 [1, 2]. 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. Current algorithms are solely based on the similarity between the text of web pages (e.g., Altavista) and/or the link structure (e.g., Google). With the advent of the semantic web an increasing amount of web pages include semantic annotations. Wouldn't it make sense to exploit this structured information for ranking? If yes, then we would need similarity measures for these semantic annotations.

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, and/or automatic database schema matching rely on good similarity measures. Thus, similarity prediction algorithms exploiting ontology structures should be a central element in the semantic web researcher's toolbox. Most semantic-web systems, however, use traditional logic approaches where corresponding objects are determined by perfect matches (or subsumption) and similarity 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 concepts (complex objects) in an ontology. In particular, it experimentally compares a number of adapted or existing computational measures assembled in a similarity framework with the human judgment of the similarity of objects in an ontology. The contributions of this paper are the following. First, it presents a framework of similarity measures adapted to the use in ontologies of concepts called SimPack. The framework structure shows how to map the similarity of concepts in ontologies to other, well researched similarity approaches, laying the basis for the adoption of additional measures. Second, it provides an extensive empirical evaluation of the implemented measures against a similarity "gold standard" established in an experiment with 94 subjects spanning in two different ontologies. Third, the experimental results reconfirm theories from psychology on varying nature of human similarity assessment. Last, it introduces two (varyingly accurate) methods to predict a person's similarity assessment style, paving the way to a personalized similarity measure.

The paper is structured as follows: Next, we review the literature on object similarity and present our similarity framework called SimPack, which is implemented as a generic Java class library. Then, we provide a detailed explanation of our experimental setup, present the results of the experiments, and discuss limitations of the presented study. We close with a discussion of related work and some ideas for future work.

2 Similarities in Ontologies

The question of similarity is a heavily researched subject in the computer science, artificial intelligence, psychology, and linguistics literature. Typically, those studies focus on the similarity between vectors [1, 3], strings [4], trees or graphs [5], or simple objects [6]. In our case we are interested in the similarity between concepts (complex objects) in ontologies. To clarify our discussion of similarity measures, we will first introduce a formal framework of concepts in ontologies.

2.1 Formal Framework

In this section we introduce our formal framework. Note, that it puts only very few constraints on the underlying ontology to facilitate the application of our similarity measures to a variety of different ontology formalisms (supporting monotonic inheritance such as description logic [7], or non-monotonic cases such as Courteous Logic Programs [8] or Flora [9]). Thus, our framework doesn't strive to be complete, but defines a minimal set of features.

Defintion 1: An ontology O consists of a set of concepts $\mathbb{C}_O$, a set of individuals $\mathbb{I}_O$ as well as a set of properties $\mathbb{P}_O$. In OWL [10] or RDFS [11] terminology, a concept simply refers to *owl:class* or *rdfs:class*. An ontology can graphically be represented by a rooted, labelled and unordered tree.

Definition 2: A concept C_O^X (in the ontology O of the type X) may have one or more properties $\mathbb{P}_{C_O^X} = \{P^1_{C_O^X}, \cdots, P^n_{C_O^X}\}$ defined on it, where n is the number of properties.

Definition 3: Properties have values which may either be of a primitive type (datatype) like *string*, *integer* or *double*, or of a complex type $C_O^X \in \mathbb{C}_O$ (objecttype). The *domain* of a property is the concept the property is defined on whereas the *range* of a property means the type of its value.

Definition 4: A concept C_O^X is said to be a descendant of concept C_O^Y if it is a direct (or indirect) sub-concept of C_O^Y. Think of X and Y as being the "type" of the two concepts, analogous to the types in object-oriented software languages like Java for instance. Therefore concept C_O^X inherits all properties of its super-concepts C_O^Y. The inheritance relationship is denoted by $C_O^X \rightarrow C_O^Y$ which means C_O^X is a child or, in general, a descendant of C_O^Y (in description logic "speak": C_O^Y subsumes C_O^X). The inverse relationship $C_O^X \leftarrow C_O^Y$ means C_O^Y is a parent, or more generally, an ancestor of C_O^X. Note that multiple inheritance is possible for the ontologies considered in this paper.

Definition 5: The set of individuals (extensions) of concept $C_O^X \in \mathbb{C}_O$ is denoted by $\mathbb{E}_{C_O^X} = \{I^1_{C_O^X}, \cdots, I^n_{C_O^X}\}$ where n is the number of individuals of that concept. Individuals refer to class instances in modern object-oriented programming languages.

Definition 6: The **M**ost **R**ecent **C**ommon **A**ncestor $MRCA$ of two concepts C_O^X and C_O^Y is the closest subsumer of both C_O^X and C_O^Y. In graph terminology, this means the sum of the distances of C_O^X to $MRCA$ and C_O^Y to $MRCA$ is minimal.

Any concept can be mapped (with possible loss of information) to a set containing all of its properties. Given that the properties of a concept aren't ordered, its set mapping can only be transformed to a vector (or string) if a specific ordering function *ord()* for the properties is supplied. At the moment, we have defined three different mappings:

Mapping 1 (property names): The names of the properties (including inherited properties) of a concept C_O^X are stored in a vector of strings. $M_1 : V_{C_O^X} = \{Pname_1, \cdots, Pname_n\}$ where n is the number of properties of concept C_O^X.

Mapping 2 (property-range pairs): The names of the properties (including inherited properties) of a concept C_O^X are stored in a vector of strings together with the ranges of the properties. $M_1 : V_{C_O^X} = \{Pname_1, Rname_1, \cdots, Pname_n, Rname_n\}$ where n is the number of properties of concept C_O^X. If the range is of primitive type, its name is used.

Mapping 3 (all property-range pairs in preorder): Starting at concept C_O^X, the names of its properties are sorted alphabetically and its ranges processed recursively in preorder. If the range of a property is again of type *concept*, the concept's properties are processed in the same way and so forth. The resulting vector of property-range pairs represents a tree-like structure of the tree rooted at C_O^X. If the ranges of concept C_O^X are all of primitive type, the result is the same as for M_2.

We now assembled the necessary definitions and mappings for a similarity framework, which includes similarity measures for concepts in ontologies. We found most of the similarity measures elsewhere in the literature and adopted them for the use in ontologies. All measures were implemented in our Java-based generic similarity framework SimPack, which can be easily adopted to a variety of data-structures. The remainder of this section will discuss the measures. Given the space constraints it will not present the full similarity framework but will limit the discussion to the representative subset we used in the evaluation.

2.2 Ontology-based Similarity Measures

The most intuitive similarity measure of concepts in an ontology is their distance within the ontology [12, 13] defined as the number of sub-/super-concept (is-a) relationships between them. These measures make use of the hierarchical ontology structure for determining the semantic similarity between concepts. As ontologies can be represented by rooted, labelled and unordered trees where edges between concepts represent relationships, distances between concepts can be computed by counting the number of edges on the path connecting two concepts. Sparrows, for example, 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 concepts 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 is usually defined as the shortest path going through a common ancestor or as the general shortest path, potentially connecting two concepts 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 problem with this approach is (1) that it is dependend on the design of the ontology and (2) that it relies on the notion that edges in an ontology represent uniform distances, i.e. it assumes that all semantic links are of equal weight.

One possibility of determining the semantic similarity between concepts is sim_{edge} mentioned by Resnik [14] (but normalized) and is a variant of the edge-counting method converting it from a distance to a similarity measure:

$$sim_{edge}(C_O^X, C_O^Y) = \frac{2 * MAX - len(C_O^X, C_O^Y)}{2 * MAX} \tag{1}$$

Here, MAX is the length of the longest path from the root of the ontology to any of its leaf-concepts and $len(C_O^X, C_O^Y)$ is the length of the shortest path from C_O^X to C_O^Y. Another variation of the edge-counting method is the conceptual similarity measure introduced by Wu & Palmer [15]:

$$sim_{con} = \frac{2 * N_3}{N_1 + N_2 + 2 * N_3} \tag{2}$$

where N_1, N_2 are the distances from concept C_O^X, C_O^Y resp. to its most recent common ancestor $MRCA(C_O^X, C_O^Y)$. N_3 is the distance from $MRCA(C_O^X, C_O^Y)$ to the root of ontology O.

Another measure presented by Leacock & Chodorow [16] combines the edge-counting method with information content measures (see section 2.3):

$$sim_{LC}(C_O^X, C_O^Y) = -\log\left(\frac{len(C_O^X, C_O^Y)}{2 * MAX}\right) \tag{3}$$

2.3 Information-Theory-based Similarity Measures

The problem of the ontology distance is that it is highly dependent on the (oftentimes) subjective construction of the ontology. To address this problem researchers in the NLP domain have proposed measuring the similarity between two concepts (in their case words or synsets) in an ontology (i.e., WordNet [17]) in terms of information-theoretic entropy measures [14, 19]. Specifically, Lin [19] argues that a class (in his case a word) is defined by its use. The information of a class is specified as the probability of encountering a class's (or one of its descendant's) use. In cases where many instances are available, the probability P of encountering a class's use can be computed over the instance corpus. Alternatively, when the instance space is sparsely populated (as currently in most semantic web ontologies) or when instances are also added as subclasses with is-a relationships (as with some taxonomies), then we propose to use the probability of encountering a subclass of a class. The entropy of a class is the negative log of that probability. Resnik [14] defined the similarity as

$$sim_{res}(C_O^X, C_O^Y) = \max_{C_O^Z \in \mathbb{CA}(C_O^X, C_O^Y)} [-\log P_{C_O^Z}] \tag{4}$$

where $\mathbb{CA}(C_O^X, C_O^Y)$ is the set of common ancestors of C_O^X and C_O^Y. $P_{C_O^Z}$ is the probability of encountering a concept of type Z, or just the frequency of concept type Z in O.

Lin defined the similarity between two concepts slightly different as

$$sim_{lin}(C_O^X, C_O^Y) = \frac{2 * \log P_{MRCA(C_O^X, C_O^Y)}}{\log P_{C_O^X} + \log P_{C_O^Y}} \tag{5}$$

Intuitively, this measure specifies similarity as the probabilistic degree of Overlap of descendants between two concepts. Modeling his evaluation on an experiment by Miller and Charles [23], which uses human subjects to rate the similarity between 30 noun pairs, Resnik also shows that this information-theory-based method provides significant improvement (correlation 0.79) over traditional edge methods (correlation 0.60).

2.4 Vector-Space and String-based Similarity Measures

Vector-Space-based Similarity Measures

A whole group of similarity measures operate on vectors of equal length. To simplify their discussion, we will discuss all measures as the similarity between the vectors $\mathbf{x}$ and $\mathbf{y}$, which are assumed to be computed from the concepts C_O^X and C_O^Y using the mappings $M_{1,2,3} : C_O^X \rightarrow \mathbf{x}$ outlined in section 2.1. We will also assume that all vectors have been (1) scaled to the same length by adding zeros for properties (or property-range pairs) that aren't included in the other vector and (2) ordered in the exact same way. The result is a vector space where each object occupies a distinct position. The typically used similarity measures for vectors are the cosine measure, the extended Jaccard measure and the overlap measure.

$$sim_{cosine}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}||_2 \cdot ||\mathbf{y}||_2} \tag{6}$$

$$sim_{jaccard}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}||_2^2 + ||\mathbf{y}||_2^2 - \mathbf{x} \cdot \mathbf{y}} \tag{7}$$

$$sim_{overlap}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\min(||\mathbf{x}||_2^2, ||\mathbf{y}||_2^2)} \tag{8}$$

Here, $||\mathbf{x}||$ denotes the L^1-norm of $\mathbf{x}$, i.e. $||\mathbf{x}|| = \sum_{i=1}^{n} |x_i|$, whereas $||\mathbf{x}||_2$ is the L^2-norm, thus $||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$. The cosine measure quantifies the similarity between two vectors as the cosine of the angel between the two vectors whereas the extended Jaccard measure computes the similarity as the ratio of the number of shared attributes to the number of common attributes [18]. In addition, the Dice measure [19] and the Euclidean distance are also implemented in our framework:

$$sim_{dice}(\mathbf{x}, \mathbf{y}) = \frac{2 \cdot \mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}||_2^2 + ||\mathbf{y}||_2^2} \tag{9}$$

$$d_{euclid}(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||_2 \tag{10}$$

The metric Euclidean distance is converted from a distance to a similarity measure using the following formula [19]:

$$sim_{dist}(\mathbf{x}, \mathbf{y}) = \frac{1}{1 + d_{dist}(\mathbf{x}, \mathbf{y})} \tag{11}$$

where $dist \in \{euclid, manhattan, \dots\}$ is one of the Minkowski distances ($p = 1$ for manhattan, $p = 2$ for euclid) given by

$$L_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{y}_i|^p \right)^{\frac{1}{p}} \tag{12}$$

String-based Similarity Measures

The similarity between strings is often described as the edit distance (also called the Levenshtein Distance [20]), the number of changes necessary to turn one string into another string. 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. Obviously, this approach can be adapted to strings of concepts (i.e., vectors of strings as the result of the mappings of section 2.1) rather than strings of characters by calculating the number of insert, remove, and replacement operations to transform one vector $\mathbf{x}$ to another vector $\mathbf{y}$, which is defined as $xform(\mathbf{x}, \mathbf{y})$. But should each type of transformation have the same weight? Isn't the replacement transformation, for example, comparable with a deleting procedure followed by an insertion procedure? Hence, we could argue that the cost function c should have the behavior $c(delete) + c(insert) \geq c(replace)$. Thus, we can calculate the worst case transformation cost $xform_{wc}(\mathbf{x}, \mathbf{y})$ of $\mathbf{x}$ to $\mathbf{y}$ replacing all concept parts of $\mathbf{x}$ with parts of $\mathbf{y}$, then deleting the remaining parts of $\mathbf{x}$, and inserting additional parts of $\mathbf{y}$. The worst case cost is then used to normalize the edit distance resulting in

$$sim_{Levenshtein}(C_O^X, C_O^Y) = \frac{xform(\mathbf{x}, \mathbf{y})}{xform_{wc}(\mathbf{x}, \mathbf{y})} \tag{13}$$

2.5 Full-text Similarity Measure

Last but not least, we decided to add a standard full-text similarity measure sim_{tfidf} to our framework. Essentially, we exported a full-text description of all concepts in an ontology to their textual representation and built an index over the descriptions using Apache Lucene [21]. Then we used a Porter Stemmer [22] to reduce all words to their stems and applied a standard, full-text tf-idf algorithm as described in [1] to compute the similarity between concepts. Tf-idf 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. In such a tf-idf scheme, the vectors of term weights can then be compared using one of the vector-space-based similarity measures as named in section 2.4.

3 Experimental Evaluation

The similarity framework introduced in section 2.1 provides a first catalog of ontology-based similarity metrics. 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

[1] Alternatively, we could have evaluated the measures in a realistic application for similarity measures. Such an analysis on an early and limited set of measures can be found in [citation suppressed].

human subjects were asked to assess the similarity between two concepts following [25, 26] who found that human judgments give the best assessments of the "goodness" of a similarity measure. 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.

3.1 Study Design

To establish our gold standard we first needed a suitable experimental setup. We found that the experiment described in [23], which relies on human judgments, has become the benchmark in determining the similarity of words in NLP research. We adapted their experimental design to the use with concepts in ontologies as follows: First, we had to find a number of suitable concept 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, we carefully evaluated the answers statistically. We will now visit each of these steps in detail.

Lord and colleagues [4] found that it was difficult to run any user-based evaluation with complicated ontologies (a gene ontology in their case). Therefore, we decided to use two different ontologies whose elements we considered to be understandable to our subjects. The first ontology is the MIT Process Handbook ontology (PH) [27], which contains over 5000 organizational processes and has been carefully developed for over 10 years. For the second ontology, we used the Suggested Upper Merged Ontology (SUMO) [28], which has been under development as an IEEE Standard. We used the PH ontology for our pilot study and the SUMO ontology to get an idea of how well our findings will generalize across other ontologies.

From the PH ontology we selected 40 and from the SUMO ontology 13 concept pairs 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 an object and its specialization.
- One object was paired with itself.

Each pair was then either turned into a part of into a web-page (in the PH case) or paper survey (in the SUMO case) using an on-line survey tool (see figure 1). In the web experiment the subjects were asked to assess the similarity between two processes on a scale from 1 (no similarity) to 5 (identical). We also inquired 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 objects the similarity was observed by the subjects – a notion that similarity researchers in the social sciences have found to be central [6]. 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. When finishing either surveys, the subjects where 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.

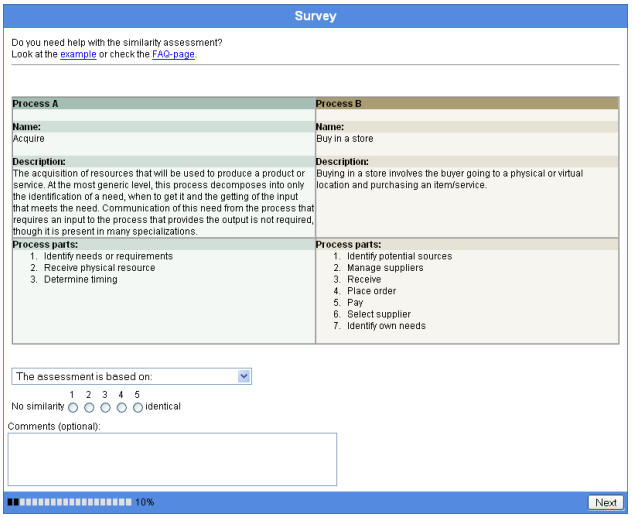


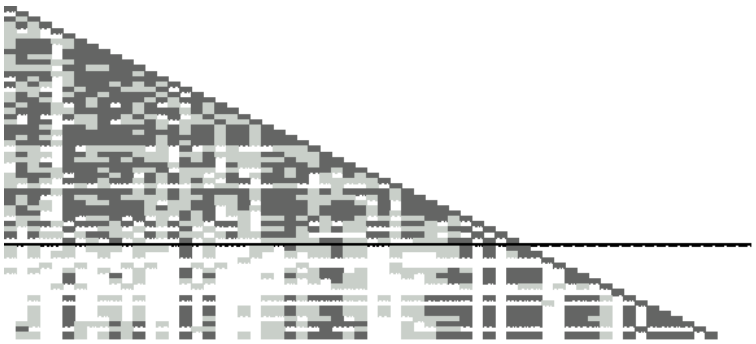
Figure 1: Sample survey page showing two objects to compare.

To avoid skewed results due to prior knowledge we deliberately recruited our subjects from three different groups: computer scientists, students, and linguists. From each group an approximately equal number of subjects participated totaling in 94 survey participants, larger than any other study we found in the literature and promising highly significant results. Each subject either took the paper (i.e., SUMO) or web (i.e., PH) survey.

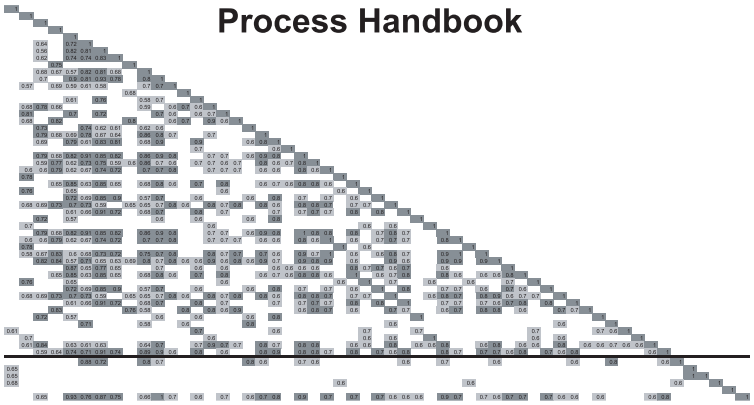
3.2 First Results/Data Analysis

To assess the quality of the similarity measures we compared their assessments with those of the subjects. We used the corrected Spearman correlation coefficient rs [29], which compares bindings (corrected ranks) of assessments because (1) the measures provided nominal predictions, the subjects' assessments where on an ordinal scale, and (2) the prediction of some measures was non-linear complicating their comparison using traditional correlations. The coefficient rs is between -1 and 1, where 1 represents perfectly correlated sets, -1 inversely correlated sets, and 0 completely uncorrelated sets. Values of $rs \geq 0.56$, respectively $rs \leq -0.56$, are taken as some correlation (i.e., correlating with $\alpha \geq 5\%$, where $n = 13$) and values of $rs \geq 0.7$, respectively $rs \leq -0.7$, as good correlations (i.e., $\alpha \geq 0.1\%$). Specifically, we took each series of similarity assessment (by either human subjects or measures) and compared it to every other assessment of the same ontology using the corrected Spearman rank correlation.

The resulting correlations for the ontologies are shown in figure 2. At a first glance the result looks rather mediocre. A large part of the measures don't even seem to correlate with a lower probability. After careful consideration, however, we find the following interesting results in the



(a) Sumo heatmap



(b) PH heatmap

Figure 2: Greyscale representation of Spearman correlations between subjects and algorithms (light grey: $|rs| \geq 0.56$, dark grey: $|rs| \geq 0.7$; black line separates subjects from measures)

data. Consider, for example, the PH ontology, the measures 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 similarity measure with the human subjects' assessments (as shown by a t-test at levels below 0.005 for all but one measure; below 0.9 for the information theory measure). Also, one can clearly see that some subjects' assessments are very different from others. In the SUMO ontology, most subjects are well correlated among each other and with a few of the measures. Also, we see that no similarity measure performs consistently well across all ontologies, supporting presumptions in the literature that similarity measures might not perform steadily across ontologies [25].

It seems as if there are distinctively different prediction approaches to predict accuracy, which seem to lead to different assessment types. Indeed, it turns out that the measures and the subject's assessments are grouped rather well into clusters (Rayleigh quotient; SUMO: 1.58; PH: 0.22). For the SUMO ontology, the edge and information-based (without sim_{con}), sequence-based, and the vector-based (with sim_{con} and sim_{tfidf}) measures each result in a coherent cluster. With a few exceptions the vector-based cluster correlates highest with the subjects. In the PH ontology, the within subjects cohesion is much lower. But again they can be clustered into 2 cohesive clusters with the measures, the first consisting of the vector and sequence-based methods (again with sim_{tfidf}), the second containing the information-theoretic-based similarity measures.

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 outcome also provides rationale for the recent surge of personalized web search services by companies such as Google™and Eurekster™. Note, that our finding contradicts findings from the NLP domain, where information-theoretic-based measures consistently outperform the ontology-based measures [14, 19, 25, 30, 31]. The reason for this difference might lie in the underlying ontologies. WordNet is essentially an ontology of "simple" objects while we explicitly investigated ontologies of complex objects.

Indeed, it turns out that the clusters are grouped either around the composition of the concepts or its ontological nature. In both ontologies we found that the vector-based measures correlated well with one group of people, reflecting assessments using the composition of the concepts, or the ontology, reflecting the conceptual nature of the class-relationships. This finding is, again, highly, consistent with studies in the psychology literature [6], which states that people base their similarity assessments on different features of the concepts to be compared – and such features can either be of compositional or conceptual (i.e., ontological) nature.

In this section we presented our adaptation of the well established semantic similarity experiment by Miller and Charles [23]. We asked 94 subjects from three different populations to assess the similarity between carefully chosen process (concept) pairs from within two ontologies. When comparing the assessed similarities using the corrected Spearman's rank correlation we found that (1) subjects seemed to use different assessment approaches, (2) the measures and human assessments could be grouped into cohesive clusters, and (3) we saw that the similarity measures could reflect the different features of the processes (concepts) used as a basis for the similarity assessments. Last, we raised the question whether the applicability of a similarity measure is ontology dependent, as some measures unexpectedly outperformed others in different ontologies, also contradicting findings with WordNet.

4 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 success in itself it would be desirable to find a way to predict a person's cluster membership in order to use a similarity measure best suited to his/her personal similarity assessment style. Given the nature of the clusters hypothesized above one obvious approach would be to use a subject's self-reported assessment method and potentially demographic information to make such a prediction. Another approach would be to learn a subject's assessment method (and, thus, cluster membership) from the assessments themselves.

For both approaches we decided to establish the "true" cluster of a subject. We used k-means clustering to form "objective" clusters based on the data. The features used for clustering where the subject's (and similarity measure's) Spearman's correlation with all other measures (i.e., the vectors would correspond to the rows of the similarity matrix, half of which is represented in figure 2). Also, given that we didn't know the number of clusters, we decided to run all tests in the section

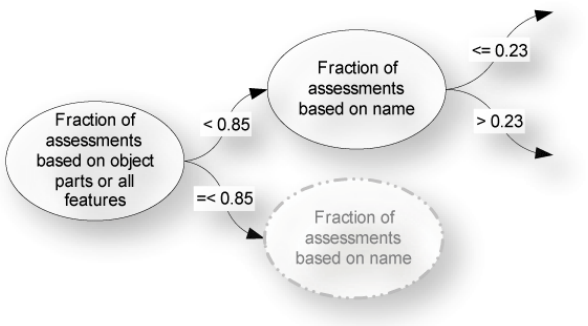


Figure 3: Decision tree model resulting from the 4 cluster SUMO case (pruned/partial view).

with k = 2, 3, and 4. We chose 4 as a maximum to reflect the four groups of similarity measures established in the similarity framework.

4.1 Personalization Using Self-reported Data

To predict a subject's cluster-membership based on her/his self-reported data we used the off-the-shelf decision tree learner J48, the Weka [32] implementation of C4.5 machine learning algorithm [33]. As an input to the algorithm we used the subjects' self-reported explanations of how they performed the similarity assessments. We complemented this information with the (rather limited) demographic information gathered at the end of the survey. We then evaluated the quality of the measures' 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.

When applied to the SUMO ontology, J48 had a 46% accuracy when predicting the cluster membership for the 4-cluster case. The resulting decision tree (see figure 3) ignores all demographic information and relies exclusively on the self-reported assessment strategy. It first considers the number of assessments that relied both on the concepts' parts and on the combination of all features. On the next level it distinguishes the assessments based on the name. Unfortunately, given the uniformity of the answers for the SUMO ontology, the models for the 3-cluster and 2-cluster case degenerated to an assignment of the majority class. Within the PH ontology the algorithm provided a 70% accuracy (when predicting the cluster membership for the two class problem and three class problem). Again, 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.

While the accuracies for the two models are definitely not too good they are within the range of the acceptable. More importantly, the resulting model reconfirms our hypothesis about the basis for cluster membership. When trying to apply this approach practically it has the drawback that it relies on the users' self-report of the similarity measure. As such it requires users to declare how they make an assessment before being able to assign them to their possible cluster. This is impractical as the question in itself might (1) bias the users, (2) might be too abstract when asked independent of examples, and (3) users might change their preferences over time without declaring so. As a consequence, we find this approach to be limited of usefulness in practice.

4.2 Personalization Based on Use

Avoiding the need for self-reported data altogether one could assign users a suitable similarity measure by observing their behavior. This also has the advantage that it doesn't bias the user by asking any questions about how they made their choices and might capture changing preferences. We investigated one approaches to personalization based on use, which we call the sub-sampling approach.

The sub-sampling approach bases on the assumption that only a part of a user's similarity assessments are sufficient to predict his/her overall behavior. As a consequence one could assume that the similarity measure best correlating with only a part of his/her similarity assessments should predict the cluster of the similarity measure best correlating with the rest of his/her predictions. The experimental procedure to test this approach was complicated. Using a 2-fold cross validation we divided the data into the data used for prediction and the data used for testing. From the prediction data we chose n similarity assessments and tested whether the similarity measure algorithm best correlating with it was in the same cluster as the algorithm best predicting the test data. Figure 4 graphs the prediction accuracy when changing the number of initial samples n for the SUMO ontology (note that we averaged the results for all possible combinatorial choices of n and display the resulting variance with the error bars). The figure shows how this approach is clearly successful in predicting the cluster membership of a user.[2]

In the last section we found that people's similarity assessments could be clearly grouped into cohesive clusters. We even hypothesized about the nature of the clusters. This section set out to predict a user's cluster membership in order to facilitate a personalized choice of a suitable similarity measure. We investigated two approaches, one based on self-reported data; another based only on observed similarity assessments by the user. We found that the prediction based on self-reported data had an acceptable accuracy, but was limited due to its need for and the accuracy of those self-reports – a requirement deemed to be impractical. Nonetheless, the resulting models seemed to reconfirm our (Genter and colleagues'[6]) hypothesis that the clustering occurred due to the different features of the concepts (processes) taken into account for the similarity assessments. We also found that the observation-based sub-sampling approach cluster predictions provided a better accuracy (in particular, in the two cluster case).

Summarizing we can say that it seems that the cluster membership of a user can be predicted by observing him/her. This finding could be used in applications to predict which similarity measure is most suitable to model a person's similarity assessments and, thus, provide the basis for an automatically-derived, personalized and highly accurate similarity assessment measure for concepts in ontologies – the goal we set ourselves at the onset of this paper.

5 Limitations and Future Work

The primary limitation of our evaluation lies in the type of evaluation we performed for the cluster prediction approaches. As we didn't foresee the need to evaluate a personalization approach we didn't collect enough data to cleanly split the learning from the test approaches. Luckily, we could approach a clean split in the sub-sampling approach by splitting along questions rather than people. Nonetheless, we believe that a thorough evaluation with users would be needed to reconfirm

[2]Note, that results for the PH ontology are similar and are, therefore, omitted.

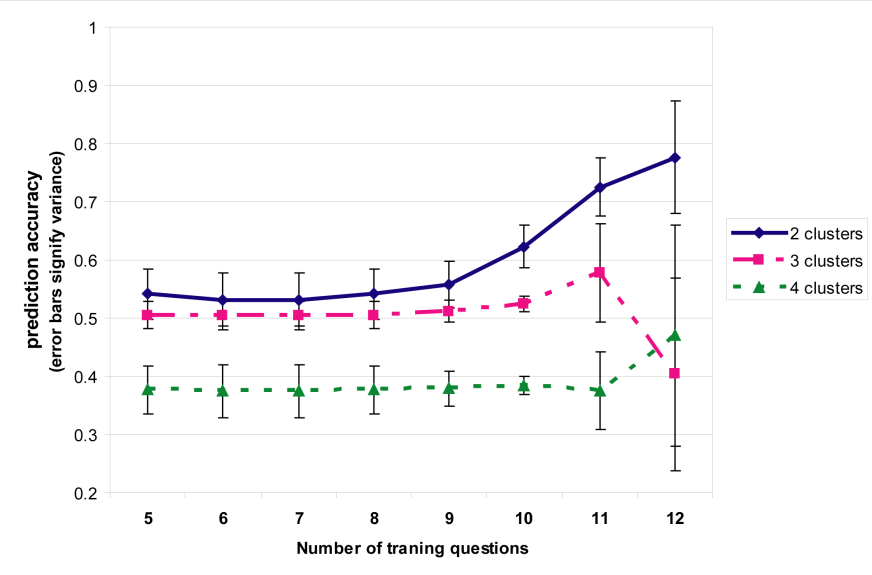


Figure 4: Learning curves for the SUMO ontology

our findings independent of our investigation. Second, our choice of ontologies was clearly limited by their availability[3]. To run statistically valid evaluations the community seems to be in dire need of more ontologies with large numbers of classes and instances, which are understandable by non-specialized human subjects. Third, we completely omitted any discussion of combined or aggregated similarity measures due to both space considerations and the even larger number of subjects as well as experimental data needed for a clean evaluation – a task which we intend to undertake in a future study. Next, a few subjects reported that they changed their assessment method during the test. Thus, any personalized similarity method has to take changes in the human preferences into account (i.e., concept drift). 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? One could easily imagine enhancing the proposed approaches to adapt to these changing assessment methods. Additional empirical studies are, however, needed to determine their applicability to our specific problem. Last, our approach only takes weak ontology features (such as subsumption and instantiation) into consideration. Note, that this is actually intended, as it allows to map apply our framework not only to description logic based ontologies but also to non-monotonic ones (like the process handbook).

In addition to addressing the limitations above (apart from ”only” using the weak ontology features), we also intend to apply our measures for cross-ontology comparisons. This is a particularilly interesting application field, as it addresses one of the major problems of the semantic web: the integration (or mapping) between ontologies of different origin. Furthermore, we are in the process of implementing a series of tree and graph based similarity measures for in SimPack. These measures are typically computationally very expensive (and might, therefore, be unsuitable for some applications)

[3]We also looked into using the OpenCyC ontology (see `http://www.opencyc.org/`). It did, however, have the disadvantage that none of the properties had a domain or a range specified, making the use of the vector and string based similarity measures impossible.

6 Related Work

The NLP literature provides the largest group of related work. Motivated by Resnik's studies [14, 34], a number of papers describe improvements to his information-theoretic measure. Wu and Palmer [15] focus on the semantic representation of verbs in computer systems and find those measures well applicable in machine translation. Jiang and Conrath [31] 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 [25] support our claim that the suitability of similarity measures might be 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 [30] 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 [19] 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. While outperforming Resnik's similarity algorithm slightly it does still require a probabilistic model of the application domain. This limitation makes it problematic for smaller ontologies.

Di Noia et al. [35] 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 correspondenc..." to the average human subject's assessment and refer to ongoing large-scale experiments for further details. Similarly, Stojanovic et al. [36] present a search engine, which significantly improves its ranking using an ontology-based inferencing process. Both investigations differ from ours in the focus on ranking retrieved objects rather than similarity measures in general. They are complementary to our study, in that a personalized similarity measure like the ones we present might be able to improve retrieval results even further, as found in information retrieval projects [37]. Using an experiment with 37 subjects Rodriguez and Egenhofer [38] 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 also shows the potential that similarity measures have for supporting translations between ontologies. Focusing on the bioinformatics application domain, Lord et al. [4] 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 [39] 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 [40]. They don't report any evaluation of their approach. Andreasen et al. [41] 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 catalog of similarity measures using a similar-size subject pool across multiple ontologies 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 doesn't use any complex (or aggregate) objects.

7 Conclusions

In this paper we argued that similarity measures in ontologies, a central component of techniques such a clustering, data-mining, semantic sense disambiguation, ontology translations, automatic database schema matching, and simple object comparison, deserve more attention. We assembled a catalog of algorithms and compared them with an experimentally derived gold standard, which we obtained by surveying 94 human subjects in two different ontologies. We found that human predictions varied in some ontologies, 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 two personalization approaches that predict a subject's cluster membership providing surprisingly accurate similarity assessments for the subjects in our study. We found that the algorithm basing on sampling had a superior predication accuracy than the one based on self-reported data.

This study provides a first investigation of personalized similarity measures in ontologies. Finding such measures is an important task for a variety of algorithms with wide applicability for the semantic web and the web in general. 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.

8 Acknowledgements

The authors would like to thank the MIT Process Handbook as well as IEEE and Adam Pease for making available the data on which the evaluation is based. This work was partially supported by the Swiss National Science Foundation grant 200021-100149/1.

Bibliography

[1] R. Baeza-Yates and B. d. A. Ribeiro-Neto. Modern Information Retrieval. ACM Press, 1999.

[2] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, vol. 30, pp. 107-117, 1998.

[3] G. Salton and M. J. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[4] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating Semantic Similarity Measures Across the Gene Ontology: The Relationship Between Sequence and Annotation. Bioinformatics, vol. 19, pp. 1275-1283, 2003.

[5] D. Shasha and K. Zhang. Approximate Tree Pattern Matching. In Pattern Matching Algorithms, A. Apostolico and Z. Galil, Eds., pp. 341-371. Oxford University Press, 1997.

[6] D. Gentner and J. Medina. Similarity and the Development of Rules. Cognition, vol. 65, pp. 263-297, 1998.

[7] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider. The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, 2003.

[8] B. Grosof, Y. Larbou, and H. Chan. A Declarative Approach to Business Rules in Contracts: Courteous Logic Programs in XML. Proceedings of the 1st ACM Conference on Electronic Commerce (EC'99), Denver, CO, Nov 3-5, 1999.

[9] G. Yang and M. Kifer. Well-Founded Optimism: Inheritance in Frame-Based Knowledge Bases. Intl. Conference on Ontologies, DataBases, and Applications of Semantics for Large Scale Information Systems (ODBASE'02), October 2002.

[10] F. Van Harmelen, J. Hendler, I. Horrocks, D. McGuinness, P. Patel-Schneider, and L. A. Stein. OWL Web Ontology Language Reference. World Wide Web Consortium, Cambridge, MA, W3C Working Draft WD-owl-ref-20030331, 31 March 2003.

[11] RDF Vocabulary Description Language 1.0: RDF Schema, `http://www.w3.org/TR/2004/REC-rdf-schema-20040210/`

[12] J. H. Lee, M. H. Kim, and Y. J. Lee. Information Retrieval Based on Conceptual Distance in IS-A Hierarchies. Journal of Documentation, vol. 49, pp. 188-207, 1993.

[13] R. Rada, H. Mili, E. Bicknell, and M. Bletner. Development and Application of a Metric on Semantic Nets. IEEE Transactions on Systems, Man, and Cybernetics, vol. 19, pp. 17-30, 1989.

[14] P. Resnik. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research, vol. 11, pp. 95-130, 1999.

[15] Z. Wu and M. Palmer. Verb Semantics and Lexical Selection. 32nd Annual Meeting of the Associations for Computational Linguistics, Las Cruces, New Mexico, 1994.

[16] C. Leacock and M. Chodorow. Combining Local Context and WordNet Similarity for Word Sense Identification. In WordNet: An Electronic Lexical Database, C. Fellbaum, MIT Press, 1998.

[17] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to WordNet: An On-line Lexical Database. Cognitive Science Laboratory, Princeton University, Princeton, Technical Report 1993.

[18] Alexander Strehl and Joydeep Ghosh and Raymond Mooney. Impact of Similarity Measures on Web-page Clustering. Proceedings of the 17th National Conference on Artificial Intelligence: Workshop of Artificial Intelligence for Web Search (AAAI'00), 30-31 July 2000, Austin, Texas, USA. 2000.

[19] D. Lin. An Information-Theoretic Definition of Similarity. 15th International Conference on Machine Learning (ICML'98), Madison, WI, 1998.

[20] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady, vol. 10, pp. 707-710, 1966.

[21] Apache Lucene, `http://lucene.apache.org/java/docs/`

[22] Snowball: Quick Introduction, `http://snowball.tartarus.org/`

[23] G. A. Miller and W. G. Charles. Contextual Correlates of Semantic Similarity. Language and Cognitive Processes, vol. 6, pp. 1-28, 1991.

[24] Michael Lee and Brandon Pincombe and Matthew Welsh. A Comparison of Machine Measures of Text Document Similarity with Human Judgments. Submitted manuscript.

[25] A. Budanitsky and G. Hirst. Semantic Distance in WordNet: An Experimental, Application-Oriented Evaluation of Five Measures. 2nd meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'01), Pittsburgh, PA, 2001.

[26] S. Blok, D. Medin, and D. Osherson. Probability from Similarity. AAAI Conference on Commonsense Reasoning, Stanford, CA, 2002.

[27] T. W. Malone, K. Crowston, J. Lee, B. Pentland, C. Dellarocas, G. Wyner, J. Quimby, C. Osborn, A. Bernstein, G. Herman, M. Klein, and E. O'Donnell. Tools for Inventing Organizations: Toward a Handbook of Organizational Processes. Management Science, vol. 45, pp. 425-443, 1999.

[28] I. Niles and A. Pease. Towards a Standard Upper Ontology. International Conference on Formal Ontologies in Information Systems (FOIS'01), Ogunquit, ME, 2001.

[29] L. Sachs. Angewandte Statistik. 10 ed. Springer, 2002.

[30] M. Jarmasz and S. Szpakowicz. Roget's Thesaurus and Semantic Similarity. University of Ottawa, School of Information Technology and Engineering, Technical Reports TR-2003-01, 2001.

[31] J. J. Jiang and D. W. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. International Conference on Research on Computational Linguistics (ROCLING X), Taiwan, 1997.

[32] I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufman Publishers, 2000.

[33] J. R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

[34] P. Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. 14th International Joint Conference on Artificial Intelligence, Montreal, 1995.

[35] T. Di Noia, E. Di Sciascio, F. M. Donini, and M. Mongiello. A System for Principled Matchmaking in an Electronic Marketplace. 12th International World Wide Web Conference, Budapest, Hungary, 2003.

[36] N. Stojanovic, R. Studer, and L. Stojanovic. An Approach for the Ranking of Query Results in the Semantic Web. 2nd International Semantic Web Conference (ISWC'03), Sanibel Island, FL, USA, 2003.

[37] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive Web Search Based on User Profile Constructed without any Effort from Users. 13th International World Wide Web Conference (WWW'04), New York, NY, 2004.

[38] M. A. Rodriguez and M. J. Egenhofer. Determining Semantic Similarity Among Entity Classes from Different Ontologies. IEEE Transactions on Knowledge and Data Engineering, vol. 15, pp. 442-456, 2003.

[39] M. Ouzzani and A. Bouguettaya. Efficient Access to Web Services. IEEE Xplore: Internet Computing, vol. 8, No. 2, pp. 34-44, 2004.

[40] M. Klein and A. Bernstein. Towards High-Precision Service Retrieval, Proceedings of the First International Semantic Web Conference on The Semantic (ISWC'02), pp. 84-101, 2002.

[41] T. Andreasen, H. Bulskov, and R. Knappe, "From Ontology over Similarity to Query Evaluation," presented at 2nd CoLogNET-ElsNET Symposium - Questions and Answers: Theoretical and Applied Perspectives, Amsterdam, Holland, 2003.